

## ON STOCHASTIC PROXIMAL GRADIENT ALGORITHMS

YVES F. ATCHADÉ, GERSENDE FORT, AND ERIC MOULINES

**ABSTRACT.** We study a perturbed version of the proximal gradient algorithm for which the gradient is not known in closed form and should be approximated. We address the convergence and derive a non-asymptotic bound on the convergence rate for the perturbed proximal gradient, a perturbed averaged version of the proximal gradient algorithm and a perturbed version of the fast iterative shrinkage-thresholding (FISTA) of Beck and Teboulle (2009). When the approximation is achieved by using Monte Carlo methods, we derive conditions involving the Monte Carlo batch-size and the step-size of the algorithm under which convergence is guaranteed. In particular, we show that the Monte Carlo approximations of some averaged proximal gradient algorithms and a Monte Carlo approximation of FISTA achieve the same convergence rates as their deterministic counterparts. To illustrate, we apply the algorithms to high-dimensional generalized linear mixed models using  $\ell_1$ -penalization.

**Keywords** Proximal Gradient Methods; Stochastic Optimization; Monte Carlo approximations; Non convex optimization; Perturbed Majorization-Minimization algorithms

## 1. INTRODUCTION

This paper deals with statistical optimization problems of the form:

$$(\mathbf{P}) \quad \min_{\theta \in \Theta} F(\theta) \quad \text{with } F = f + g ,$$

where  $f$  is a smooth function and  $g$  is a possibly non-smooth convex penalty term. It focuses on the case where  $f + g$  and  $\nabla f$  are both intractable.

In penalized likelihood estimation,  $f = -\ell$  where  $\ell$  is the log-likelihood function. Intractability arises because the likelihood is known only up to a normalization factor, depending on the parameter to estimate, which cannot be computed explicitly. This is the case when considering for example parameter inference in random fields or undirected graphical models. In such case,  $\ell$  and  $\nabla \ell$  are integrals with respect to (w.r.t.) some Gibbs probability measure  $\pi_\theta$  on some measurable space  $\mathbf{X}$  and known only up to the partition function  $Z_\theta$  (normalization constant). Intractable likelihood

---

2000 *Mathematics Subject Classification.* 60F15, 60G42.

Y. F. Atchadé: University of Michigan, 1085 South University, Ann Arbor, 48109, MI, United States. *E-mail address:* yvesa@umich.edu.

Gersende Fort: LTCI, CNRS & Telecom ParisTech, 46, rue Barrault 75634 Paris Cedex 13, France. *E-mail address:* gersende.fort@telecom-paristech.fr.

Eric Moulines: LTCI, Telecom ParisTech & CNRS, 46, rue Barrault 75634 Paris Cedex 13, France. *E-mail address:* eric.moulines@telecom-paristech.fr.

functions and intractable gradient also arise when dealing with hierarchical latent variable models, including missing data or mixed effects models. Here again,  $\ell(\theta)$  and  $\nabla\ell(\theta)$  are integrals w.r.t. a distribution  $\pi_\theta$  which represents the conditional distribution of the latent/missing variables given the parameter and the data. In both cases,  $\pi_\theta$  not only depends on  $\theta$ , but is often difficult to simulate and typically requires to use Markov Chain Monte Carlo (MCMC) methods.

Another source of intractability arises when doing learning on a huge data set. In this case,  $f$  is written as  $f = \sum_{i=1}^N f_i$  where  $N$  is the sample size. In this case,  $f$  and  $\nabla f$  can be computed explicitly but it is more practical to reduce the computational cost by using Monte Carlo estimation of these quantities (see e.g. the survey on incremental stochastic gradient and subgradient methods by Bertsekas (2012))

Intractable problems also occur in online learning and stochastic approximation, where the function  $f$  takes the form  $f(\theta) = \int \bar{f}(\theta; x)\pi(dx)$  with an unknown distribution  $\pi$ : the user is only provided with random samples from  $\pi$  by an oracle. The goal is to minimize some parameter-dependent functional of the distribution to minimize an empirical risk or maximize a likelihood function.

To cope with problems where  $f + g$  is intractable and possibly non-smooth, various methods have been proposed. Some of these work focused on stochastic sub-gradient and mirror descent algorithms; see Nemirovski et al. (2008); Duchi et al. (2011); Cotter et al. (2011); Lan (2012); Juditsky and Nemirovski (2012a,b). Other authors have proposed algorithms based on proximal operators to better exploit the smoothness of  $f$  and the properties of  $g$  (Hu et al. (2009); Xiao (2010)). Proximal algorithms are well established optimization algorithms for dealing with non-smooth objective functions; Beck and Teboulle (2010); Parikh and Boyd (2013); Juditsky and Nemirovski (2012a,b).

The current paper focuses on the proximal gradient algorithm (see e.g. Nesterov (2004)) and some perturbed versions: the gradient  $\nabla f(\theta_n)$  at the current estimate  $\theta_n$  is replaced by an approximation  $H_{n+1}$  to solve the problem (P). We provide sufficient conditions on the perturbation  $\eta_{n+1} = H_{n+1} - \nabla f(\theta_n)$  to obtain convergence of the sequence  $\{F(\theta_n), n \in \mathbb{N}\}$  and convergence of the sequence  $\{\theta_n, n \in \mathbb{N}\}$ . These results are obtained under weak conditions on  $f$  and we provide a characterization of the limit points of the sequences  $\{F(\theta_n), n \in \mathbb{N}\}$  and  $\{\theta_n, n \in \mathbb{N}\}$  (see Theorem 6).

In the case where  $f$  is convex, the limit set of  $\{\theta_n, n \in \mathbb{N}\}$  is the set of the solutions of (P) (see Theorem 7). Our results therefore cover the case of strongly convex function  $f$ , despite in this setting, the tools are significantly different and the rates of convergence as well (see e.g. Rosasco et al. (2014); Nitanda (2014); Xiao and Zhang (2014)). Corollary 8 gives a special emphasis to the case the approximation error  $\eta_{n+1}$  is random and  $f$  is convex. Conditions (expressed in terms of the conditional bias and the conditional variance of the error  $\eta_{n+1}$ ) are given in order to ensure convergence; these conditions improve on the earlier works by Combettes and Wajs (2005); Rosasco et al. (2014).

We then consider an averaging scheme of the perturbed proximal gradient algorithm: given non-negative weights  $\{a_n, n \in \mathbb{N}\}$  with partial sum  $A_n \stackrel{\text{def}}{=} \sum_{k=1}^n a_k$ , Theorem 10 addresses the rate of convergence of  $A_n^{-1} \sum_{k=1}^n a_k F(\theta_k)$  to the minimum

$F_*$  of the function  $F$ . Rates of convergence in expectation are also provided in the case of a stochastic perturbation  $\eta_{n+1}$  (see Corollary 11). We also extend the analysis to perturbed version of the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle (2009). Theorem 13 provides sufficient conditions for the convergence of  $\{F(\theta_n), n \in \mathbb{N}\}$ . Here again, an emphasis is given on the stochastic setting: Corollary 14 and Corollary 15 provide rates of convergence of  $\{F(\theta_n), n \in \mathbb{N}\}$  to the minimum  $F_*$  of the function  $F$ , for resp. almost-sure convergence and convergence in expectation. In this stochastic setting, our results for the averaging scheme and the perturbed FISTA improve on previous work (see e.g. Schmidt et al. (2011)).

Both the perturbed averaged proximal gradient algorithm and the perturbed FISTA depend on design parameters: the stepsize sequence  $\{\gamma_n, n \in \mathbb{N}\}$  in the proximal operator, the weight sequence  $\{a_n, n \in \mathbb{N}\}$  in the averaged proximal gradient (resp. the weight sequence in FISTA). We discuss the choice of these parameters in the case of Monte Carlo proximal gradient where at each iteration the current gradient  $\nabla f(\theta_n)$  is approximated by the mean  $m_{n+1}^{-1} \sum_{k=1}^{m_{n+1}} H_{\theta_n}(X_{n+1,k})$ . The choice of the Monte Carlo batch size  $m_{n+1}$  is also discussed under assumptions on the convergence rates to zero of the error  $\eta_{n+1}$  which are satisfied with Importance Sampling Monte Carlo and Markov chain Monte Carlo approximations.

We show that this perturbed averaged proximal gradient algorithm (resp. the perturbed FISTA) can reach the same convergence rate as the non perturbed ones for convenient choices of  $\{\gamma_n, n \in \mathbb{N}\}$ ,  $\{m_n, n \in \mathbb{N}\}$  and of the weight sequences: the perturbed averaged proximal gradient algorithm (resp. perturbed FISTA) converges at the rate  $n^{-1}$  (resp.  $n^{-2}$ ) when the stepsize is constant  $\gamma_n = \gamma$  and the Monte Carlo batch size  $\{m_n, n \in \mathbb{N}\}$  increases to infinity at a convenient rate.

However, since these algorithms require the approximation of the gradient at increasing precision (which is costly), the convergence rates expressed as a function of the number of iterations do not yield the complete picture as they do not account for the computational cost of approximating the gradient. A better measure of the performance of these algorithms is the total number of Monte Carlo samples needed to approximate the solution at a given precision  $\delta > 0$ . Our results imply that the number of Monte Carlo samples needed to achieve a precision  $\delta$  is  $O(\delta^{-2})$  for the stochastic averaged gradient proximal algorithm and  $O(\delta^{-2-\kappa})$  for some  $\kappa > 0$ , for the perturbed FISTA.

We illustrate these results with the problem of estimating the parameters of a  $\ell^1$ -penalized random effect logistic regression model. The simulation results support the theoretical findings reported above.

The paper is organized as follows. In section 2 our basic assumptions are stated, the proximal gradient algorithm is described and its main properties are recalled. In section 3, the convergence results for the perturbed proximal gradient, some averaged version and the FISTA algorithm are stated. section 4 is devoted to the case  $\eta_{n+1}$  is a Monte Carlo sum. An application of the stochastic proximal gradient algorithm to

the  $\ell^1$ -penalized inference of a logistic regression with mixed effects is presented in section 5. The proofs are postponed to section 6.

## 2. THE PROXIMAL GRADIENT ALGORITHM

In the sequel  $\Theta$  denotes a finite-dimensional Euclidean space with norm  $\|\cdot\|$  and inner product  $\langle \cdot, \cdot \rangle$ . Let  $f : \Theta \rightarrow \mathbb{R}$ , and  $g : \Theta \rightarrow (-\infty, +\infty]$  be functions. We consider the problem of finding solutions to (P) when  $f, g$  satisfy the conditions

**H1.**  $g : \Theta \rightarrow (-\infty, +\infty]$  is convex, not identically  $+\infty$ , and lower semi-continuous. The function  $f : \Theta \rightarrow \mathbb{R}$  is continuously differentiable and there exists a finite constant  $L$  such that, for all  $\theta, \theta' \in \Theta$ ,

$$\|\nabla f(\theta) - \nabla f(\theta')\| \leq L\|\theta - \theta'\| ,$$

where  $\nabla f$  denotes the gradient of  $f$ .

This assumption implicitly holds in all the sequel. Note that since the function  $g$  is convex, it is continuous on its domain  $\text{Dom}(g) = \{\theta \in \Theta, |g(\theta)| < \infty\}$ . In the examples below,  $f$  is the negative log-likelihood (and is denoted by  $-\ell$ ) and  $g$  is a penalty function.

**Example 1** (High-dimensional logistic regression with random effects). We model binary responses  $\{Y_i\}_{i=1}^N \in \{0, 1\}$  as  $N$  conditionally independent realizations of a random effect logistic regression model,

$$Y_i | \mathbf{U} \stackrel{\text{ind.}}{\sim} \text{Ber}(s(x'_i \beta + \sigma z'_i \mathbf{U})) , \quad 1 \leq i \leq N , \quad (1)$$

where  $x_i \in \mathbb{R}^p$  is the vector of covariates,  $z_i \in \mathbb{R}^q$  are (known) loading vector,  $\text{Ber}(\alpha)$  denotes the Bernoulli distribution with parameter  $\alpha \in [0, 1]$ ,  $s(x) = e^x / (1 + e^x)$  is the cumulative distribution function of the standard logistic distribution. The random effect  $\mathbf{U}$  is assumed to be standard Gaussian  $\mathbf{U} \sim N_q(0, I)$ .

The log-likelihood of the observations at  $\theta = (\beta, \sigma) \in \Theta = \mathbb{R}^p \times (0, \infty)$  is given by

$$\ell(\theta) = \log \int \prod_{i=1}^N s(x'_i \beta + \sigma z'_i \mathbf{u})^{Y_i} (1 - s(x'_i \beta + \sigma z'_i \mathbf{u}))^{1-Y_i} \phi(\mathbf{u}) d\mathbf{u} , \quad (2)$$

where  $\phi$  is the density of a  $\mathbb{R}^q$ -valued standard Gaussian random vector. The number of covariates  $p$  is possibly larger than  $N$ , but only a very small number of these covariates are relevant which suggests to use the elastic-net penalty

$$\lambda \left( \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) , \quad (3)$$

where  $\lambda > 0$  is the regularization parameter,  $\|\beta\|_r = (\sum_{i=1}^p |\beta_i|^r)^{1/r}$  and  $\alpha \in [0, 1]$  controls the trade-off between the  $\ell^1$  and the  $\ell^2$  penalties. In this example,

$$g(\theta) = \lambda \left( \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right) . \quad (4)$$

Define the conditional log-likelihood of  $\mathbf{Y} = (Y_1, \dots, Y_N)$  given  $\mathbf{U}$  (the dependence upon  $\mathbf{Y}$  is omitted) by

$$\ell_c(\theta|\mathbf{u}) = \sum_{i=1}^N \left\{ Y_i (x'_i \beta + \sigma z'_i \mathbf{u}) - \ln(1 + \exp(x'_i \beta + \sigma z'_i \mathbf{u})) \right\},$$

and the conditional distribution of the random effect  $\mathbf{U}$  given the observations  $\mathbf{Y}$  and the parameter  $\theta$

$$\pi_\theta(\mathbf{u}) = \exp(\ell_c(\theta|\mathbf{u}) - \ell(\theta)) \phi(\mathbf{u}). \quad (5)$$

The Fisher identity implies that the gradient of the log-likelihood (2) is given by

$$\nabla \ell(\theta) = \int \nabla_\theta \ell_c(\theta|\mathbf{u}) \pi_\theta(\mathbf{u}) \, d\mathbf{u} = \int \left\{ \sum_{i=1}^N (Y_i - s(x'_i \beta + \sigma z'_i \mathbf{u})) \begin{bmatrix} x_i \\ z'_i \mathbf{u} \end{bmatrix} \right\} \pi_\theta(\mathbf{u}) \, d\mathbf{u}.$$

The Hessian of the log-likelihood  $\ell$  is given by (see e.g. (McLachlan and Krishnan, 2008, Chapter 3))

$$\nabla^2 \ell(\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta^2 \ell_c(\theta|\mathbf{U})] + \text{Cov}_{\pi_\theta} (\nabla_\theta \ell_c(\theta|\mathbf{U}))$$

where  $\mathbb{E}_{\pi_\theta}$  and  $\text{Cov}_{\pi_\theta}$  denotes the expectation and the covariance with respect to the distribution  $\pi_\theta$ , respectively. Since

$$\nabla_\theta^2 \ell_c(\theta|\mathbf{u}) = - \sum_{i=1}^N s(x'_i \beta + \sigma z'_i \mathbf{u}) (1 - s(x'_i \beta + \sigma z'_i \mathbf{u})) \begin{bmatrix} x_i \\ z'_i \mathbf{u} \end{bmatrix} \begin{bmatrix} x_i \\ z'_i \mathbf{u} \end{bmatrix}',$$

and  $\sup_{\theta \in \Theta} \int \|\mathbf{u}\|^2 \pi_\theta(\mathbf{u}) \, d\mathbf{u} < \infty$  (see Appendix A),  $\nabla^2 \ell(\theta)$  is bounded on  $\Theta$ . Hence,  $\nabla \ell(\theta)$  satisfies the Lipschitz condition showing that H1 is satisfied.  $\square$

**Example 2** (Network structure estimation). Consider a Markov random field over a finite set  $\mathbf{X}$  with joint probability distribution

$$f_\theta(x_1, \dots, x_p) = \frac{1}{Z_\theta} \exp \left\{ \sum_{i=1}^p \theta_{ii} B_0(x_i) + \sum_{1 \leq j < i \leq p} \theta_{ij} B(x_i, x_j) \right\}, \quad (6)$$

where  $B_0 : \mathbf{X} \rightarrow \mathbb{R}$  and  $B : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$  is a symmetric function and  $Z_\theta$  is the normalizing constant, which cannot (in general) be computed explicitly. The real-valued symmetric  $p \times p$  matrix  $\theta$  defines the network structure which is the parameter of interest. We consider the problem of estimating  $\theta$  from  $N$  realizations  $\{x^{(i)}\}_{i=1}^N$  from (6) where  $x^{(i)} = (x_1^{(i)}, \dots, x_p^{(i)}) \in \mathbf{X}^p$  under sparsity assumptions.

Let  $\Theta$  denote the space of  $p \times p$  symmetric matrices equipped with the (modified) Frobenius inner product

$$\langle \theta, \vartheta \rangle \stackrel{\text{def}}{=} \sum_{1 \leq k \leq j \leq p} \theta_{jk} \vartheta_{jk}, \text{ with norm } \|\theta\|_2 \stackrel{\text{def}}{=} \sqrt{\langle \theta, \theta \rangle}.$$

We compute the penalized maximum likelihood estimate over  $\mathcal{K}_a$  with  $\ell^1$ -penalty, where  $\mathcal{K}_a \stackrel{\text{def}}{=} \{\theta \in \Theta : 0 \leq \theta_{ij} \leq a, \text{ for all } i \neq j\}$  for some constant  $a > 0$ . Notice

that we do not penalize the diagonal terms of  $\theta$ . Therefore, we compute the minimum of  $F = -\ell + g$  where

$$\ell(\theta) = \frac{1}{N} \sum_{i=1}^N \langle \theta, \bar{B}(x^{(i)}) \rangle - \log Z_\theta \text{ and } g(\theta) = \lambda \sum_{1 \leq k < j \leq p} |\theta_{jk}| + \mathbb{I}_{\mathcal{K}_a}(\theta) ;$$

the matrix-valued function  $\bar{B} : \mathbf{X}^p \rightarrow \mathbb{R}^{p \times p}$  is defined by

$$\bar{B}_{ii}(x) = B_0(x_i) \quad \bar{B}_{ij}(x) = B(x_i, x_j), i \neq j ,$$

$\lambda$  is the positive regularization parameter and  $\mathbb{I}_{\mathcal{C}}$  is defined by

$$\mathbb{I}_{\mathcal{C}}(\vartheta) = \begin{cases} 0 & \text{if } \vartheta \in \mathcal{C} , \\ +\infty & \text{otherwise .} \end{cases} \quad (7)$$

Observe that  $g$  is convex, not identically equal to  $+\infty$  and is lower-continuous (since  $\mathcal{C}$  is closed, see (Bauschke and Combettes, 2011, Example 1.25)). Upon noting that (6) is a canonical exponential model, (Shao, 2003, Section 4.4.2) shows that  $\theta \mapsto -\ell(\theta)$  is convex and

$$\nabla \ell(\theta) = \frac{1}{N} \sum_{i=1}^N \bar{B}(x^{(i)}) - \int_{\mathbf{X}^p} \bar{B}(z) f_\theta(z) \mu(dz) , \quad (8)$$

where  $\mu$  is the counting measure on  $\mathbf{X}^p$ . In addition, (see Appendix B)

$$\|\nabla \ell(\theta) - \nabla \ell(\vartheta)\|_2 \leq p \left( (p-1) \text{osc}^2(B) + \text{osc}^2(B_0) \right) \|\theta - \vartheta\|_2, \quad (9)$$

where for a function  $\tilde{B} : \mathbf{X} \times \mathbf{X} \rightarrow \mathbb{R}$ ,  $\text{osc}(\tilde{B}) = \sup_{x,y,u,v \in \mathbf{X}} |\tilde{B}(x,y) - \tilde{B}(u,v)|$ . Hence, H1 holds. □

**Example 3** (large scale convex optimization). Consider the case when  $f$  is the average of many smooth component functions  $f_i$  i.e.

$$f(\theta) = \frac{1}{N} \sum_{i=1}^N f_i(\theta) ,$$

and all the component functions  $f_i$  admit a Lipschitz-continuous gradient which can be explicitly computed. Such a situation occurs in statistical inference problems (in that case  $N$  is the total number of observations and  $f_i$  is the loss function associated to the  $i$ -th observation), in distributed optimization, in the minimization of an expected value when the random variable takes many finite values; see e.g. Bertsekas (2012) and the references therein.

In the large scale problems i.e. when  $N \gg 1$ , the computational cost of  $\nabla f(\theta)$  is so high that it is advocated to use incremental methods by substituting, at iteration  $n+1$ , the gradient  $\nabla f(\theta_n)$  by  $m_{n+1}^{-1} \sum_{k=1}^{m_{n+1}} \nabla f_{I_k}(\theta_n)$  where  $\{I_k, k \geq 1\}$  are, conditionally to the past  $\mathcal{F}_n$ , i.i.d. random variables drawn at random in  $\{1, \dots, N\}$  and independent of  $\mathcal{F}_n$ . □

The proximal map (see e.g. Parikh and Boyd (2013)) associated to  $g$  is defined for  $\gamma > 0$  by:

$$\text{Prox}_\gamma(\theta) \stackrel{\text{def}}{=} \text{Argmin}_{\vartheta \in \Theta} \left\{ g(\vartheta) + \frac{1}{2\gamma} \|\vartheta - \theta\|^2 \right\}. \quad (10)$$

Note that under H 1, there exists a unique point  $\vartheta$  minimizing the RHS for any  $\theta \in \Theta$  and  $\gamma > 0$ . There are many important examples of penalty functions for which the proximal map is tractable. For example, for the projection (7), the proximal operator is the orthogonal projection on  $\mathcal{C}$ , that is the map  $\Pi_{\mathcal{C}} : \Theta \rightarrow \Theta$  such that  $\|\theta - \Pi_{\mathcal{C}}(\theta)\| = \min_{\vartheta \in \mathcal{C}} \|\theta - \vartheta\|$ . For the elastic-net penalty (3),  $\text{Prox}_\gamma(\theta)$  is the component-wise soft-thresholding operator defined as

$$(\mathbf{s}_{\gamma, \lambda, \alpha}(\theta))_i = \begin{cases} \frac{\theta_i - \gamma\lambda\alpha}{1 + \gamma\lambda(1 - \alpha)} & \text{if } \theta_i \geq \gamma\lambda\alpha, \\ \frac{\theta_i + \gamma\lambda\alpha}{1 + \gamma\lambda(1 - \alpha)} & \text{if } \theta_i \leq -\gamma\lambda\alpha, \\ 0 & \text{if } \theta_i \in (-\gamma\lambda\alpha, \gamma\lambda\alpha). \end{cases} \quad (11)$$

Note that the case  $\alpha = 1$  corresponds to the Lasso penalty and the case  $\alpha = 0$  corresponds to the ridge-regression penalty. When both of these penalties are combined, and  $\mathcal{C}$  is a rectangle, the proximal operator is  $\Pi_{\mathcal{C}}(\mathbf{s}_{\gamma, \lambda, \alpha}(\theta))$ .

We can now introduce the proximal gradient algorithm (see Beck and Teboulle (2009)):

**Algorithm 1** (Proximal gradient algorithm). Let  $\theta_0 \in \text{Dom}(g)$  denote the starting estimate, and  $\{\gamma_n, n \in \mathbb{N}\}$  be a sequence of positive step sizes. Given  $\theta_n$ , compute

$$\theta_{n+1} = \text{Prox}_{\gamma_{n+1}}(\theta_n - \gamma_{n+1} \nabla f(\theta_n)). \quad (12)$$

The proximal gradient is a special instance of the Majorization-Minimization (MM) technique (see (Beck and Teboulle, 2009, Section 1.3.)). Consider the surrogate function  $\vartheta \mapsto Q_\gamma(\vartheta|\theta)$  given by

$$\begin{aligned} Q_\gamma(\vartheta|\theta) &\stackrel{\text{def}}{=} f(\theta) + \langle \nabla f(\theta), \vartheta - \theta \rangle + \frac{1}{2\gamma} \|\vartheta - \theta\|^2 + g(\vartheta) \\ &= f(\theta) + \frac{1}{2\gamma} \|\vartheta - (\theta - \gamma \nabla f(\theta))\|^2 - \frac{\gamma}{2} \|\nabla f(\theta)\|^2 + g(\vartheta), \end{aligned} \quad (13)$$

for a positive constant  $\gamma \in (0, 1/L]$ , where  $L$  is defined in H 1. By construction,  $F(\theta) = Q_\gamma(\theta|\theta)$  for all  $\theta \in \Theta$  and  $F(\vartheta) \leq Q_\gamma(\vartheta|\theta)$  for all  $\theta, \vartheta \in \Theta$ . Hence,  $\vartheta \mapsto Q_\gamma(\vartheta|\theta)$  is a *majorizing function* for  $\vartheta \mapsto F(\vartheta)$  which suggests to iteratively compute the minimum of the majorizing surrogate, leading to (12). Observe that (12) is of the form  $\theta_{n+1} = T_{\gamma_{n+1}}(\theta_n)$  with the point-to-point map  $T_\gamma$  defined by

$$T_\gamma(\theta) \stackrel{\text{def}}{=} \text{Prox}_\gamma(\theta - \gamma \nabla f(\theta)) = \text{argmin}_{\vartheta \in \text{Dom}(g)} Q_\gamma(\vartheta|\theta). \quad (14)$$

Hence, when  $\gamma_n = \gamma$  for any  $n$ , any convergent sequence  $\{\theta_n, n \in \mathbb{N}\}$  has a limiting point which is a fixed point of  $T_\gamma$ . Set

$$\mathcal{L} \stackrel{\text{def}}{=} \{\theta \in \text{Dom}(g) : 0 \in f(\theta) + \partial g(\theta)\}, \quad (15)$$

where  $\partial g(\theta)$  denotes the subdifferential of the convex function  $g$  at  $\theta$  in  $\text{Dom}(g)$ . Then for any  $\gamma > 0$ ,

$$\mathcal{L} = \{\theta : \theta = T_\gamma(\theta)\} \quad (16)$$

(see Beck and Teboulle (2009); see also Lemma 20 in section 6). If in addition  $f$  is convex, then  $\mathcal{L}$  is the set of the minimizers of  $F = f + g$ :

$$\mathcal{L} = \text{Argmin}_{\theta \in \Theta} F(\theta) . \quad (17)$$

Convergence results for Algorithm 1 applied with constant stepsize ( $\gamma_n = \gamma$  for any  $n$ ) can be obtained by applying general results for monotonic optimization algorithm adapted to MM algorithms; see (Meyer, 1976, Theorem 3.1.) and (Schifano et al., 2010, Appendix A). The following convergence result is established in Section 6.2.

**Theorem 4.** *Assume H1 and set  $\gamma \in (0, 1/L]$ . Let  $\{\theta_n, n \in \mathbb{N}\}$  be the sequence given by Algorithm 1 applied with  $\gamma_n = \gamma$  for any  $n \geq 1$  and assume that the sequence  $\{\theta_n, n \in \mathbb{N}\}$  remains in a compact set  $\mathcal{K}$ . Then,*

- (i) *the set  $\mathcal{L}$  given by (15) is not empty and all accumulation points of  $\{\theta_n, n \in \mathbb{N}\}$  are in  $\mathcal{L} \cap \mathcal{K}$ ,*
- (ii) *there exists  $\theta_\star \in \mathcal{L} \cap \mathcal{K}$  such that  $\lim_n F(\theta_n) = F(\theta_\star)$ ,*
- (iii)  *$\lim_n \|\theta_{n+1} - \theta_n\| = 0$ .*

The key properties for the proof of this result is that (i)  $T_\gamma$  is monotonic (see Lemma 20 in section 6): for any  $\theta$ ,  $F(T_\gamma(\theta)) \leq F(\theta)$ , which implies that the sequence  $\{F(\theta_n), n \in \mathbb{N}\}$  is non-increasing; and (ii)  $T_\gamma : \Theta \rightarrow \Theta$  is Lipschitz (see Lemma 18 in section 6): there exists  $C$  such that for any  $\theta, \theta'$ ,  $\|T_\gamma(\theta) - T_\gamma(\theta')\| \leq C\|\theta - \theta'\|$ .

Since  $T_\gamma$  is monotonic,  $F(\theta_{n+1}) \leq F(\theta_n)$  for any  $n$  which implies that  $\{\theta_n, n \in \mathbb{N}\}$  remains in a compact set as soon as the level set  $\{\theta : F(\theta) \leq F(\theta_0)\}$  is compact.

The property (iii) implies that either the sequence  $\{\theta_n, n \in \mathbb{N}\}$  converges to  $\theta_\star \in \mathcal{L} \cap \mathcal{K}$ , or the set of limit points of this sequence forms a continuum (and the sequence fails to converge). It also implies that if  $\{\theta_n, n \in \mathbb{N}\}$  has an isolated accumulation point  $\theta_\star$ , then  $\lim_n \theta_n = \theta_\star$ . By the properties (ii) and (iii), it is easily seen that if the number of points of  $\mathcal{L}$  having a given value of  $F$  is finite, then the sequence  $\{\theta_n, n \in \mathbb{N}\}$  converges to one of these points.

We stress that this convergence result is obtained without assuming that  $f$  is convex. If the following assumption holds

**H2.** *The function  $f$  is convex and the set  $\text{argmin}_{\theta \in \Theta} F(\theta)$  is not empty*

then Theorem 4 implies the stronger convergence result (see e.g. (Beck and Teboulle, 2009, Theorem 1.2.) and references therein).

**Corollary 5.** *Assume H1, H2 and fix  $\gamma \in (0, 1/L]$ . Let  $\{\theta_n, n \in \mathbb{N}\}$  be the sequence given by Algorithm 1 applied with  $\gamma_n = \gamma$  for any  $n \geq 1$ . Then there exists  $\theta_\star \in \mathcal{L}$  such that  $\lim_n \theta_n = \theta_\star$ .*

Corollary 5 follows from Theorem 4 upon noting that in the convex case,  $\mathcal{L}$  is not empty if and only if the sequence  $\{\theta_n, n \in \mathbb{N}\}$  remains in a compact set (see Section 6.3).



## 3. PERTURBED PROXIMAL GRADIENT ALGORITHMS

As shown in the above examples, the gradient  $\nabla f$  is not always tractable so the proximal gradient does not apply. We introduce a perturbed proximal gradient algorithm whereby the gradient at the  $n$ -th iteration of the algorithm  $\nabla f(\theta_n)$  is approximated by  $H_{n+1} \in \Theta$ . This yields the algorithm

**Algorithm 2** (Perturbed Proximal Gradient algorithm). Let  $\theta_0 \in \text{Dom}(g)$  be the initial solution and  $\{\gamma_n, n \in \mathbb{N}\}$  be a sequence of positive non-increasing step sizes. For  $n \geq 1$ , given  $(\theta_1, \dots, \theta_n)$ :

- (1) Obtain  $H_{n+1} \in \Theta$ , an approximation of  $\nabla f(\theta_n)$ .
- (2) Compute  $\theta_{n+1} = \text{Prox}_{\gamma_{n+1}}(\theta_n - \gamma_{n+1}H_{n+1})$ .

The convergence of this algorithm to the solutions of the problem (P) is established when  $f$  is convex in (Combettes and Wajs, 2005, Theorem 3.4). We provide in Theorem 6 a convergence result, which holds when  $f$  is not convex and under weaker conditions on the *approximation error*:

$$\eta_{n+1} \stackrel{\text{def}}{=} H_{n+1} - \nabla f(\theta_n). \quad (18)$$

We also give sufficient conditions for the convergence to a point  $\theta_\star$  in  $\mathcal{L}$  of this algorithm with general step size, under the stronger assumption that  $f$  is convex (see Theorem 7 and Corollary 8)). We then provide in Theorem 10 convergence rates for (possibly weighted) means of the sequence  $\{F(\theta_n), n \in \mathbb{N}\}$  in the convex case. Finally, we study in Theorem 13 the rate of convergence of a perturbed version of the Fast Iterative Shrinkage-Thresholding algorithm (FISTA).

Some results are stated in the general case of a random approximation  $H_{n+1}$ , which covers the case of a deterministic approximation.

**Theorem 6.** *Assume H 1. Set  $\gamma \in (0, 1/L]$  and let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 2 with  $\gamma_n = \gamma$ . Assume in addition that the set  $\mathcal{L}$  given by (15) is not empty,  $\limsup_n \|\theta_n\|$  is finite and  $\lim_n \eta_n = 0$  a.s. Then the sequence  $\{F(\theta_n), n \in \mathbb{N}\}$  converges a.s. to a connected component of  $F(\mathcal{L})$ . If in addition,  $F(\mathcal{L})$  has an empty interior, there exists  $\theta_\star \in \mathcal{L}$  such that*

- (i)  $\lim_n F(\theta_n) = F(\theta_\star)$  a.s.
- (ii) the sequence  $\{\theta_n, n \in \mathbb{N}\}$  converges to the set  $\mathcal{L} \cap \{\theta : F(\theta) = F(\theta_\star)\}$  a.s.

*Proof.* See Section 6.4. □

By using the Lipschitz properties of  $\text{Prox}$  and  $T_\gamma$  (see Lemma 16 and Lemma 18 in section 6), it is easy to prove that for any  $\theta_\star \in \mathcal{L}$ ,  $\|\theta_{n+1} - \theta_\star\| \leq \gamma_{n+1}\|\eta_{n+1}\| + \|\theta_n - \theta_\star\|$ . Hence, by a trivial induction we have for any  $n \geq m \geq 0$ ,

$$\|\theta_{n+1} - \theta_\star\| \leq \sum_{k \geq m} \gamma_{k+1}\|\eta_{k+1}\| + \|\theta_m - \theta_\star\|. \quad (19)$$

The above inequality shows that when  $\sum_n \gamma_{n+1}\|\eta_{n+1}\| < \infty$  a.s., then  $\limsup_n \|\theta_n\| < \infty$  a.s. In addition, if the sequence  $\{\theta_n, n \in \mathbb{N}\}$  possesses a limit point  $\theta_\star$  in  $\mathcal{L}$ , then the sequence converges to  $\theta_\star$ . Nevertheless this condition on the convergence of the

sum reveals to be strong in many situations as shown in Remark 3 below (see also the case of Monte Carlo Proximal Gradient algorithms in section 4).

When  $\limsup_n \|\theta_n\|$  is finite almost-surely, the almost-sure convergence to zero of the sequence  $\{\eta_n, n \geq 1\}$  is equivalent to the almost-sure convergence to zero of the sequence  $\{\eta_n \mathbb{1}_{\|\theta_{n-1}\| \leq B}, n \geq 1\}$ , for any  $B > 0$ . Therefore, it is easily seen by application of the conditional Borel-Cantelli lemma that a sufficient condition for the almost-sure convergence to zero of  $\{\eta_{n+1} \mathbb{1}_{\|\theta_n\| \leq B}, n \geq 1\}$  is: for any  $\varepsilon > 0$ ,

$$\sum_{n \geq 1} \mathbb{P}(\|\eta_{n+1}\| \geq \varepsilon | \mathcal{F}_n) \mathbb{1}_{\|\theta_n\| \leq B} < \infty, \quad (20)$$

where  $\{\mathcal{F}_n, n \in \mathbb{N}\}$  is the filtration defined by

$$\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(\theta_0, H_1, \dots, H_n). \quad (21)$$

In section 4, the case where  $\nabla f$  is an expectation and  $H_n$  is a Monte Carlo approximation is discussed and sufficient conditions for (20) to hold are given.

When  $f$  is convex, we know from (17) that  $\mathcal{L}$  is the set of the minimizers of  $F$ ; hence,  $F(\mathcal{L})$  has an empty interior since it is a singleton of  $\mathbb{R}$ . In this case, Theorem 6 establishes the convergence of the perturbed proximal gradient sequence  $\{\theta_n, n \in \mathbb{N}\}$  to a subset of the solutions of (P). When this set is a singleton  $\{\theta_\star\}$  - which occurs for example when  $F$  is strictly convex -, the sequence  $\{\theta_n, n \in \mathbb{N}\}$  converges almost-surely to  $\theta_\star$ . This convergence result can be compared to the convergence result given by (Combettes and Wajs, 2005, Theorem 3.4.) in the convex case and when  $H_{n+1}$  is deterministic: our theorem requires a stronger condition on the stepsize  $\gamma$  but a weaker condition on the noise  $\eta_n$  (in Combettes and Wajs (2005), it is assumed that  $\gamma \in (0, 2/L)$  and  $\sum_n \|\eta_n\| < \infty$ ).

**Example** (large scale convex optimization (continued)). In this example,

$$\eta_{n+1} = \frac{1}{m_{n+1}} \sum_{k=1}^{m_{n+1}} \{\nabla f_{I_k}(\theta_n) - \nabla f(\theta_n)\}.$$

Since  $\mathbb{E}[\nabla f_{I_k}(\theta_n) | \mathcal{F}_n] = N^{-1} \sum_{i=1}^N \nabla f_i(\theta_n)$ , conditionally to the past  $\mathcal{F}_n$  the error  $\eta_{n+1}$  is unbiased:  $\mathbb{E}[\eta_{n+1} | \mathcal{F}_n] = 0$ . Let  $B > 0$  be fixed; upon noting that for any  $i \in \{1, \dots, N\}$ ,  $\theta \mapsto \nabla f_i(\theta) - \nabla f(\theta)$  is Lipschitz-continuous and  $\sup_{\|\theta\| \leq B} \|\nabla f_i(\theta)\| < \infty$ , the uniform law of large numbers implies (see e.g. (Jennrich, 1969, Theorem 2))

$$\lim_{n \rightarrow \infty} \sup_{\{\theta, \|\theta\| \leq B\}} \left| \frac{1}{n} \sum_{k=1}^n \{\nabla f_{I_k}(\theta) - \nabla f(\theta)\} \right| = 0 \quad \text{a.s.}$$

Therefore, a sufficient condition for the almost-sure convergence of  $\{\eta_n, n \in \mathbb{N}\}$  to zero is  $\lim_n m_n = +\infty$ . □

The result below addresses the boundedness and the convergence of the sequence  $\{\theta_n, n \in \mathbb{N}\}$  to a point  $\theta_\star \in \mathcal{L}$  in the convex case.

**Theorem 7.** *Assume H1 and H2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 2 with stepsizes satisfying  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ .*

(i) For any  $\theta_*$  in  $\mathcal{L}$  (see (17)) and for any  $n \geq m \geq 0$ ,

$$\begin{aligned} & \|\theta_{n+1} - \theta_*\|^2 \\ & \leq \|\theta_m - \theta_*\|^2 - 2 \sum_{k=m}^n \gamma_{k+1} \langle T_{\gamma_{k+1}}(\theta_k) - \theta_*, \eta_{k+1} \rangle + 2 \sum_{k=m}^n \gamma_{k+1}^2 \|\eta_{k+1}\|^2. \end{aligned} \quad (22)$$

(ii) Assume that for any  $\theta_* \in \mathcal{L}$ ,

$$\sum_{n \geq 0} \gamma_{n+1} \langle T_{\gamma_{n+1}}(\theta_n) - \theta_*, \eta_{n+1} \rangle \text{ exists,} \quad \sum_{n \geq 0} \gamma_{n+1}^2 \|\eta_{n+1}\|^2 < \infty. \quad (23)$$

Then, for any  $\theta_*$  in  $\mathcal{L}$ ,  $\lim_n \|\theta_n - \theta_*\|$  exists. If in addition  $\sum_n \gamma_n = +\infty$ , then there exists  $\theta_\infty \in \mathcal{L}$  such that  $\lim_n \theta_n = \theta_\infty$ .

*Proof.* See Section 6.5. □

When  $H_{n+1}$  is a random approximation, we provide below sufficient conditions for the conclusions of Theorem 7 to hold a.s. Define

$$\epsilon_n^{(1)} \stackrel{\text{def}}{=} \|\mathbb{E}[\eta_{n+1} | \mathcal{F}_n]\|, \quad \epsilon_n^{(2)} \stackrel{\text{def}}{=} \mathbb{E}[\|\eta_{n+1}\|^2 | \mathcal{F}_n]. \quad (24)$$

**Corollary 8** (of Theorem 7). *Assume H 1 and H 2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 2 with stepsizes satisfying  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ . Assume also that  $\limsup_n \|\theta_n\| < \infty$  a.s. and for any  $B \geq 0$ ,*

$$\sum_{n \geq 0} \gamma_{n+1} \left\{ \epsilon_n^{(1)} + \gamma_{n+1} \epsilon_n^{(2)} \right\} \mathbb{1}_{\|\theta_n\| \leq B} < \infty \text{ a.s.} \quad (25)$$

*Then for any  $\theta_* \in \mathcal{L}$ ,  $\lim_n \|\theta_n - \theta_*\|$  exists a.s. If in addition  $\sum_n \gamma_n = +\infty$ , then there exists  $\theta_* \in \mathcal{L}$  such that  $\lim_n \theta_n = \theta_*$  a.s.*

*Proof.* See Section 6.6 □

**Remark 9.** A closely related result has been established in (Combettes and Pesquet, 2014, Theorem 2.5) in the case  $\gamma_n = \gamma$  for any  $n$ , but under stronger conditions; in particular, the authors assume that all the limit points of  $\{\theta_n, n \in \mathbb{N}\}$  are in  $\mathcal{L}$  which is not an assumption in Corollary 8.

**Example** (large scale convex optimization (continued)). We already proved that  $\epsilon_n^{(1)} = 0$ . Since conditionally to  $\mathcal{F}_n$ ,  $\{I_k, k \in \mathbb{N}\}$  are independent and independent of  $\theta_n$ , we have

$$\epsilon_n^{(2)} \mathbb{1}_{\|\theta_n\| \leq B} = m_{n+1}^{-1} \mathbb{E}[\|\nabla f_{I_1}(\theta_n) - \nabla f(\theta_n)\|^2 | \mathcal{F}_n] \mathbb{1}_{\|\theta_n\| \leq B} \leq M m_{n+1}^{-1}$$

for a positive constant  $M$ . This inequality shows that (25) holds a.s. for any  $B > 0$  if  $\sum_n \gamma_{n+1}^2 m_{n+1}^{-1} < \infty$ . □

Let  $\{a_n, n \in \mathbb{N}\}$  be a sequence of non-negative weights, and denote

$$A_n \stackrel{\text{def}}{=} \sum_{k=1}^n a_k.$$

Theorem 10 provides a control of the weighted mean  $A_n^{-1} \sum_{k=1}^n a_k F(\theta_k) - F(\theta_*)$  where  $\theta_*$  is any minimizer of  $F$ .

**Theorem 10.** *Assume H 1 and H 2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 2 with stepsizes satisfying  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ . For any non-negative sequence  $\{a_n, n \in \mathbb{N}\}$ , any minimizer  $\theta_*$  of  $F$  and any  $n \geq 1$ ,  $A_n^{-1} \sum_{k=1}^n a_k F(\theta_k) - F(\theta_*) \leq B_n$  where*

$$B_n \stackrel{\text{def}}{=} \frac{1}{2A_n} \sum_{k=2}^n \left( \frac{a_k}{\gamma_k} - \frac{a_{k-1}}{\gamma_{k-1}} \right) \|\theta_{k-1} - \theta_*\|^2 + \frac{a_1}{2\gamma_1 A_n} \|\theta_1 - \theta_*\|^2 - \frac{1}{A_n} \sum_{k=1}^n a_k \left\{ \langle T_{\gamma_k}(\theta_{k-1}) - \theta_*, \eta_k \rangle + \gamma_k \|\eta_k\|^2 \right\}, \quad (26)$$

and  $T_\gamma$  and  $\eta_n$  are resp. given by (14) and (18).

*Proof.* See Section 6.7. □

We deduce convergence rates in expectation from Theorem 10.

**Corollary 11** (of Theorem 10). *In addition to the assumptions of Theorem 10, assume that  $\{a_n/\gamma_n, n \geq 1\}$  is non-decreasing and there exists a constant  $B$  such that  $\mathbb{P}(\sup_{n \in \mathbb{N}} \|\theta_n\| \leq B) = 1$ . Then, for any minimizer  $\theta_*$  of  $F$ ,*

$$\frac{1}{A_n} \sum_{k=1}^n a_k \mathbb{E}[F(\theta_k)] - F(\theta_*) \leq \frac{B_*^2 a_n}{2\gamma_n A_n} + \frac{1}{A_n} \sum_{k=1}^n a_k \left\{ B_* \mathbb{E}[\epsilon_{k-1}^{(1)}] + \gamma_k \mathbb{E}[\epsilon_{k-1}^{(2)}] \right\}, \quad (27)$$

where  $B_* \stackrel{\text{def}}{=} B + \|\theta_*\|$ .

*Proof.* See Section 6.8 □

Theorem 10 provides a convergence rate in a quite general setting: the stepsizes  $\{\gamma_n, n \in \mathbb{N}\}$  are not necessarily constant and the weights  $\{a_n, n \in \mathbb{N}\}$  are non-negative but otherwise arbitrary. Taking  $a_j = 1$  provides a bound for the cumulative regret.

When the approximation  $H_{n+1}$  is deterministic, the weight sequence  $a_n = 1$  and the stepsize is constant  $\gamma_n = 1/L$ , (Schmidt et al., 2011, Proposition 1) provides a bound of order  $O(1/n)$  under the assumption that  $\sum_n \|\eta_{n+1}\| < \infty$ . Note that (19) and Lemma 22, the assumption  $\sum_n \|\eta_n\| < \infty$  implies that  $\lim_n \|\theta_n - \theta_*\|$  exists for any  $\theta_* \in \mathcal{L}$ . Using the inequality  $|\langle T_{1/L}(\theta_k) - \theta_*, \eta_{k+1} \rangle| \leq L^{-1} \|\theta_k - \theta_*\| \|\eta_{k+1}\|$  (see (55)), the upper bound  $B_n$  in (26) is also  $O(1/n)$ .

Theorem 10 applied with  $\eta_n = 0$  can be used to derive the rate of convergence of the weighted mean  $\{A_n^{-1} \sum_{j=1}^n a_j F(\theta_j), n \geq 1\}$  to the minimum  $F(\theta_*)$  for the (exact) proximal gradient sequence  $\{\theta_n, n \in \mathbb{N}\}$  given by Algorithm 1. This rate is  $a_n \gamma_n^{-1} A_n^{-1}$  when  $\sup_n \|\theta_n - \theta_*\| < \infty$  and the sequence  $\{a_n/\gamma_n, n \geq 1\}$  is non decreasing. For example, if the algorithm is run with  $\gamma_n \sim C_c n^{-c}$  for  $c \in [0, 1)$ , the weighted mean converges at the rate  $n^{c-1}$  for any weight sequence  $a_n \sim n^a$  with  $a > -1$ , showing that the minimal regret is achieved by taking  $c = 0$  i.e. a constant stepsize  $\gamma_n = \gamma$ .

**Remark 12.** Consider the weighted averaged sequence  $\{\bar{\theta}_n, n \in \mathbb{N}\}$  defined by

$$\bar{\theta}_n \stackrel{\text{def}}{=} \frac{1}{A_n} \sum_{k=1}^n a_k \theta_k, \quad (28)$$

for a given sequence  $\{a_n, n \in \mathbb{N}\}$  of non-negative weights. It generalizes the Polyak-Juditsky averaging in Stochastic Approximation (Polyak and Juditsky (1992)). Under H1 and H2,  $F$  is convex so that  $F(\bar{\theta}_n) \leq A_n^{-1} \sum_{k=1}^n a_k F(\theta_k)$ . Therefore, Theorem 10 also provides convergence rates for  $F(\bar{\theta}_n) - F(\theta_*)$  and Corollary 11 provides  $L^1$ -rates.

Nesterov (1983) introduced an acceleration scheme of the deterministic gradient method that is shown to converge at a rate  $O(1/n^2)$ . The algorithm was then extended in (Beck and Teboulle, 2009, Section 1.5) to the proximal gradient algorithm, yielding FISTA. We consider a perturbed version of FISTA.

Let  $\{t_n, n \in \mathbb{N}\}$  be a sequence of positive numbers and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive stepsizes with the following property

$$t_0 = 1, \quad \gamma_{n+1} t_n (t_n - 1) \leq \gamma_n t_{n-1}^2, \quad n \geq 1. \quad (29)$$

For instance, when  $\{\gamma_n, n \in \mathbb{N}\}$  is non-increasing, (29) holds for the following sequences: (i) all sequences  $t_n \propto (n + n_0)^\beta$ , with  $\beta \in (0, 1)$ , for some  $n_0 > 0$ ; (ii)  $t_n = n/2 + 1$  and (iii) the sequence defined recursively by

$$t_{n+1} = \frac{1 + \sqrt{1 + 4t_n^2}}{2}. \quad (30)$$

It is easily seen from (30) that  $t_n \sim n/2$  as  $n$  goes to infinity.

**Algorithm 3** (Perturbed FISTA). Let  $\theta_0 \in \text{Dom}(g)$ ,  $\{t_n, n \in \mathbb{N}\}$  and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive sequences satisfying (29). Compute  $\theta_1 = \text{Prox}_{\gamma_1}(\theta_0 - \gamma_1 H_1)$ , where  $H_1$  is an approximation of  $\nabla f(\theta_0)$ . For  $n \geq 1$ , given  $(\theta_0, \dots, \theta_n)$ :

(1) Compute

$$\vartheta_n = \theta_n + t_n^{-1}(t_{n-1} - 1)(\theta_n - \theta_{n-1}). \quad (31)$$

(2) Obtain  $H_{n+1}$  an approximation of  $\nabla f(\vartheta_n)$ , and set

$$\theta_{n+1} = \text{Prox}_{\gamma_{n+1}}(\vartheta_n - \gamma_{n+1} H_{n+1}). \quad (32)$$

For this algorithm, the approximation error is defined by

$$\check{\eta}_{n+1} \stackrel{\text{def}}{=} H_{n+1} - \nabla f(\vartheta_n). \quad (33)$$

**Theorem 13.** Assume H1 and H2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 3, with  $\{t_n, n \in \mathbb{N}\}$  and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive sequences satisfying (29) and  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ . Set  $\bar{\Delta}_n \stackrel{\text{def}}{=} (1 - t_n)\theta_n + t_n T_{\gamma_{n+1}}(\vartheta_n)$  where  $T_\gamma$  is given by (14).

(i) If  $\lim_n \gamma_{n+1} t_n^2 = +\infty$  and there exists  $\theta_* \in \mathcal{L}$  such that

$$\sum_{n \geq 0} \gamma_{n+1} t_n \langle \bar{\Delta}_n - \theta_*, \check{\eta}_{n+1} \rangle \text{ exists,} \quad \sum_{n \geq 0} \gamma_{n+1}^2 t_n^2 \|\check{\eta}_{n+1}\|^2 < \infty,$$

then  $\lim_n F(\theta_n) = F(\theta_*)$  and all the limit points of  $\{\theta_n, n \in \mathbb{N}\}$  are in  $\mathcal{L}$ .

(ii) For all  $n \geq 1$  and any minimizer  $\theta_*$  of  $F$ ,  $F(\theta_{n+1}) - F(\theta_*) \leq B_n t_n^{-2} \gamma_{n+1}^{-1}$  where

$$B_n \stackrel{\text{def}}{=} \gamma_1 (F(\theta_1) - F(\theta_*)) + \frac{1}{2} \|\theta_1 - \theta_*\|^2 + \sum_{k=1}^n \gamma_{k+1}^2 t_k^2 \|\check{\eta}_{k+1}\|^2 - \sum_{k=1}^n \gamma_{k+1} t_k \langle \bar{\Delta}_k - \theta_*, \check{\eta}_{k+1} \rangle .$$

(iii) If  $\sum_n \gamma_{n+1} t_n \|\check{\eta}_n\| < \infty$  then  $\sup_n \|\bar{\Delta}_n\| < \infty$ .

*Proof.* See Section 6.9.  $\square$

When  $H_{n+1}$  is random, we provide below sufficient conditions for the conclusions of Theorem 13 to hold a.s. Set  $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(\theta_0, H_1, \dots, H_n)$  and

$$\check{\epsilon}_n^{(1)} \stackrel{\text{def}}{=} \|\mathbb{E}[\check{\eta}_{n+1} | \mathcal{F}_n]\| , \quad \check{\epsilon}_n^{(2)} \stackrel{\text{def}}{=} \mathbb{E}[\|\check{\eta}_{n+1}\|^2 | \mathcal{F}_n] . \quad (34)$$

**Corollary 14** (of Theorem 13). *Assume H1 and H2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 3, with  $\{t_n, n \in \mathbb{N}\}$  and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive sequences satisfying (29) and  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ . If  $\limsup_n \|\vartheta_n\| < \infty$  a.s. ,  $\lim_n \gamma_n t_n^2 = +\infty$  and for any  $B > 0$ , there exist constants  $\{M_n, n \in \mathbb{N}\}$  such that*

$$\left( \check{\epsilon}_n^{(1)} + \check{\epsilon}_n^{(2)} \right) \mathbb{1}_{\cap_{k \leq n} \{\|\vartheta_k\| \leq B\}} \leq M_n \quad \text{a.s.} \quad \sum_{n \geq 0} \gamma_{n+1} t_n \{1 + \gamma_{n+1} t_n\} M_n < \infty , \quad (35)$$

then

- (i) Almost-surely,  $\lim_n F(\theta_n) = F(\theta_*)$  for  $\theta_* \in \mathcal{L}$  and the limit points of  $\{\theta_n, n \in \mathbb{N}\}$  are in  $\mathcal{L}$ .
- (ii) For all  $n \geq 1$  and any minimizer  $\theta_*$  of  $F$ ,  $F(\theta_{n+1}) - F(\theta_*) \leq B_n t_n^{-2} \gamma_{n+1}^{-1}$  a.s. where the r.v.  $B_n$  is given by Theorem 13 and is finite a.s.

*Proof.* See Section 6.10.  $\square$

**Example** (large scale convex optimization (continued)). Following the same lines as above, it is easily established that  $\check{\epsilon}_n^{(1)} = 0$  and for any  $B$ , there exists a constant  $M$  such that for any  $n$ ,

$$\check{\epsilon}_n^{(2)} \mathbb{1}_{\cap_{k \leq n} \{\|\vartheta_k\| \leq B\}} \leq \check{\epsilon}_n^{(2)} \mathbb{1}_{\|\vartheta_n\| \leq B} \leq M m_{n+1}^{-1} .$$

Therefore (35) holds if  $\sum_n \gamma_{n+1} t_n (1 + \gamma_{n+1} t_n) m_{n+1}^{-1} < \infty$ .  $\square$

We will show in section 4 (see (39) and (40)) that the condition (35) is satisfied when  $H_{n+1}$  is a Markov chain Monte Carlo approximation of the intractable gradient  $\nabla f(\vartheta_n)$ .

Theorem 13 provides a convergence rate of  $F(\theta_n)$  to the minimum  $F(\theta_*)$  for arbitrary sequences  $\{\gamma_n, n \in \mathbb{N}\}$  and  $\{t_n, n \in \mathbb{N}\}$ . It extends the previous work by Schmidt et al. (2011) which considers the case of a deterministic approximation  $H_{n+1}$ , a fixed step size  $\gamma_n = 1/L$  and the sequence  $t_n = n/2 + 1$ : under the assumption that  $\sum_{j \geq 1} t_j \|\check{\eta}_{j+1}\| < \infty$ , they prove that (i)  $\sup_n \|\bar{\Delta}_n - \theta_*\| < \infty$  and (ii) the rate of convergence is  $n^{-2}$ . In this specific setting, Theorem 13 shows that  $\sup_n \|\bar{\Delta}_n - \theta_*\| < \infty$  and  $\sup_n B_n < \infty$  so that the rate of convergence is  $n^{-2}$  too.

Theorem 13 applied with  $\check{\eta}_n = 0$  provides the rate of convergence of  $\{F(\theta_n), n \in \mathbb{N}\}$  to  $F(\theta_*)$  for the (non perturbed) FISTA. This rate is  $t_n^{-2}\gamma_n^{-1}$ . For example, if the algorithm is run with  $t_n \sim Cn$  and with  $\gamma_n \sim C_c n^{-c}$  for  $c \in [0, 2)$ , the exact algorithm converges at the rate  $n^{c-2}$ ; this rate is  $O(n^{-2})$  when the stepsize is constant  $\gamma_n = \gamma$  for some  $\gamma \in (0, 1/L]$  (see (Beck and Teboulle, 2009, Theorem 1.4)).

We deduce convergence rates in expectation from Theorem 13.

**Corollary 15** (of Theorem 13). *Assume H1 and H2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 3, with  $\{t_n, n \in \mathbb{N}\}$  and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive sequences satisfying (29) and  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ . If  $\mathbb{P}(\sup_n \|\theta_n\| \leq B) = 1$  and*

$$\sum_n \gamma_{n+1} t_n \|\check{\epsilon}_n^{(1)}\|_2 + \sum_n \gamma_{n+1}^2 t_n^2 \mathbb{E} [\check{\epsilon}_n^{(2)}] < \infty, \quad (36)$$

then for any minimizer  $\theta_* \in \mathcal{L}$ ,  $\mathbb{E}[F(\theta_{n+1})] - F(\theta_*) \leq \bar{B}_n t_n^{-2} \gamma_{n+1}^{-1}$  where

$$\begin{aligned} \bar{B}_n \stackrel{\text{def}}{=} \gamma_1 (\mathbb{E}[F(\theta_1)] - F(\theta_*)) + \frac{(B + \|\theta_*\|)^2}{2} + 2 \sum_{j=1}^n \gamma_{j+1}^2 t_j^2 \mathbb{E} [\check{\epsilon}_j^{(2)}] \\ + \sum_{j=1}^n \gamma_{j+1} t_j \|\check{\epsilon}_j^{(1)}\|_2 (\|\bar{\Delta}_j\|_2 + \|\theta_*\|), \end{aligned}$$

with  $\|U\|_2 \stackrel{\text{def}}{=} \sqrt{\mathbb{E}[U^2]}$ ,  $\check{\epsilon}_n^{(i)}$  is given by (34) and  $\sup_j \|\bar{\Delta}_j\|_2 < \infty$ .

*Proof.* See Section 6.10. An upper bound of  $\|\bar{\Delta}_j\|_2$  is provided in the proof.  $\square$

Note that when  $\mathbb{P}(\sup_n \|\theta_n\| \leq B) = 1$ , there exists  $B'$  such that  $\mathbb{P}(\sup_n \|\vartheta_n\| \leq B') = 1$ ; so the condition (36) is satisfied if (35) holds.

#### 4. MONTE CARLO PROXIMAL GRADIENT ALGORITHM

In this section, we consider the case

**H3.** for all  $\theta \in \Theta$ ,

$$\nabla f(\theta) = \int_{\mathbf{X}} H_\theta(x) \pi_\theta(dx), \quad (37)$$

for some probability measure  $\pi_\theta$  on a measurable space  $(\mathbf{X}, \mathcal{B})$ , and a measurable function  $H : \Theta \times \mathbf{X} \rightarrow \Theta$ ,  $(\theta, x) \mapsto H_\theta(x)$ , satisfying  $\int \pi_\theta(dx) \|H_\theta(x)\| < \infty$ .

Assumption H3 is satisfied in many problems (see Example 1 and Example 2 presented in section 2).

We derive convergence rates for the algorithms 2 and 3 when  $H_{n+1}$  is a Monte Carlo approximation of the expectation  $\nabla f(\theta_n)$ :

$$H_{n+1} = m_{n+1}^{-1} \sum_{j=1}^{m_{n+1}} H_{\theta_n}(X_{n+1,j}).$$

We refer to  $\{X_{n+1,j}\}_{j=1}^{m_{n+1}}$  as a Monte Carlo batch and  $m_n$  as the Monte Carlo batch-size. We will discuss how to choose the number of Monte Carlo samples  $m_n$ , the stepsize sequence  $\{\gamma_n, n \in \mathbb{N}\}$  and the sequence  $\{a_n, n \in \mathbb{N}\}$ .

In the naive Monte Carlo approach,  $\{X_{n+1,j}\}_{j=1}^{m_{n+1}}$  is, conditionally to the past  $\mathcal{F}_n$  (see (21)), a sequence of i.i.d. random variables distributed according to the current target distribution  $\pi_{\theta_n}$ . Then the approximation bias  $\mathbb{E}[\eta_{n+1} | \mathcal{F}_n]$  vanishes; if in addition there exists  $p \geq 2$  such that  $\sup_{\|\theta\| \leq B} \pi_{\theta} |H_{\theta}|^p < \infty$  then

$$\mathbb{E} [\mathbb{E} [\|\eta_{n+1}\|^p | \mathcal{F}_n] \mathbb{1}_{\|\theta_n\| \leq B}] \leq C m_{n+1}^{-p/2}, \quad (38)$$

for some finite constant  $C$  depending on  $B$  (see e.g. (Petrov, 1995, Chapter 2)). Therefore (20) holds if for some  $p \geq 2$ ,  $\sum_j m_{j+1}^{-p/2} < \infty$ , which is a very mild requirement.

A more flexible approach is to compute  $H_{n+1}$  using Markov chain Monte Carlo (MCMC) algorithms. Let  $P_{\theta}$  be a Markov kernel with invariant distribution  $\pi_{\theta}$ , and let  $\nu_{\theta}$  be some initial distribution on  $\mathbf{X}$ ; see e.g. Robert and Casella (2005) and the references therein. At the  $n$ -th iteration,  $\{X_{n+1,j}\}_{j \geq 0}^{m_{n+1}}$  is, conditionally to the past  $\mathcal{F}_n$ , a Markov chain with initial distribution  $\nu_{\theta_n}$  and Markov kernel  $P_{\theta_n}$ . In many instances the kernel is geometrically ergodic with ergodicity constants which are controlled uniformly over  $\theta \in \mathcal{K}$  where  $\mathcal{K}$  is a compact subset of  $\Theta$ ,

$$\sup_{\theta \in \mathcal{K}} \|\nu_{\theta} P_{\theta}^j H_{\theta} - \pi_{\theta} H_{\theta}\| \leq C_{\mathcal{K}} \rho_{\mathcal{K}}^j$$

with  $\rho_{\mathcal{K}} \in [0, 1)$  and  $C_{\mathcal{K}} < \infty$ ; see for example (Fort et al., 2011, Lemma 2.3.). In this setting, the approximation bias  $\mathbb{E}[\eta_{n+1} | \mathcal{F}_n]$  can be computed as

$$\mathbb{E}[\eta_{n+1} | \mathcal{F}_n] = m_{n+1}^{-1} \sum_{j=0}^{m_{n+1}-1} \left( \nu_{\theta_n} P_{\theta_n}^j H_{\theta_n} - \pi_{\theta_n} H_{\theta_n} \right),$$

so that

$$\|\mathbb{E}[\eta_{n+1} | \mathcal{F}_n]\| \mathbb{1}_{\mathcal{K}}(\theta_n) \leq C_{\mathcal{K}} (1 - \rho_{\mathcal{K}})^{-1} m_{n+1}^{-1}. \quad (39)$$

Furthermore, by (Fort and Moulines, 2003, Proposition 12), for all  $p \geq 2$

$$\mathbb{E} [\|\eta_{n+1}\|^p | \mathcal{F}_n] \mathbb{1}_{\mathcal{K}}(\theta_n) \leq \left( 6^p (1 + C_r) \sup_{\theta \in \mathcal{K}} \sup_{j \geq 0} \nu_{\theta} P_{\theta}^j G_{\theta} \right) m_{n+1}^{-p/2}, \quad (40)$$

where  $G_{\theta}(x) \stackrel{\text{def}}{=} \|\sum_{k \geq 0} P_{\theta}^k (H_{\theta} - \pi_{\theta} H_{\theta})(x)\|^p$  and  $C_r$  is the Burkholder-Rosenthal-Pinelis constant (see e.g. (Hall and Heyde, 1980, Theorem 2.12)). Sufficient conditions for these upper bounds to be finite can be found in (Fort et al., 2011, Lemma 2.3.). It follows again that (20) holds if there exists  $p \geq 2$  such that  $\sum_n m_{n+1}^{-p/2} < \infty$ .

Let us investigate the rates of convergence of

$$\mathcal{E}_n \stackrel{\text{def}}{=} \frac{1}{A_n} \sum_{k=1}^n a_k F(\theta_k) - F(\theta_{\star}),$$

that can be deduced from Corollary 11 as a function of the design parameters  $m_n, \gamma_n$  and  $a_n$ . To that goal, suppose that

$$\mathbb{E}[\epsilon_n^{(1)}] \leq C_1 m_{n+1}^{-1}, \quad \mathbb{E}[\epsilon_n^{(2)}] \leq C_2 m_{n+1}^{-1}, \quad \mathbb{P}(\sup_{n \in \mathbb{N}} \|\theta_n - \theta_{\star}\| \leq B) = 1, \quad (41)$$



with  $\epsilon_n^{(i)}$  given by (24),  $C_1, C_2 \geq 0$ ,  $\theta_\star$  a minimizer of  $F$  and  $B > 0$ . In addition, we choose

$$a_n = n^a, \quad m_n = \lfloor C_b n^b \rfloor, \quad \gamma_n = C_c n^{-c},$$

where  $\lfloor \cdot \rfloor$  denotes the lower integer part,  $a > -1$ ,  $b \geq 0$ ,  $c \geq -a \vee 0$  and  $C_b, C_c > 0$ .

Let us first discuss the case where the approximation of  $\nabla f(\theta_n)$  is unbiased (i.e.  $C_1 = 0$  in (41)). For a given  $c \in [0, 1]$ , the maximal rate of the RHS of (27) is  $n^{1-c}$  which is obtained by choosing  $b \geq 0$  such that  $b + c = 1 - c$ . This yields to  $c \in [0, 1/2]$ ,  $a > -c$  and  $b = 1 - 2c$ ; the rate of convergence is  $1/n^{1-c}$ . It is maximal when  $c = 0$  which requires to increase the batch size linearly  $m_n = \lfloor C_b n \rfloor$ ; the number of iterations required to obtain  $\mathcal{E}_n \leq \delta$  increases at a rate  $\delta^{-1}$  and the total number of simulations increases as  $\delta^{-2}$ . The rate  $\delta^{-2}$  can not be improved by other choices of the parameters but can be reached by other strategies: by choosing  $c \in [0, 1/2]$ ,  $a > -c$  and  $b = 1 - 2c$ , the rate is also  $\delta^{-2}$ .

Let us now discuss the case when the approximation error is biased. For a given  $c \in [0, 1]$ , the maximal rate of the RHS of (27) is  $n^{1-c}$  which is obtained by choosing  $b = 1 - c$  and  $a > 0$ . This rate is maximal when  $c = 0$ ,  $b = 1$ , which yields a convergence rate  $O(1/n)$ . Here again, for this choice, the total number of Monte Carlo samples required to obtain  $\mathcal{E}_n \leq \delta$  increases as  $\delta^{-2}$ .

As a conclusion of the above discussion, we report in Table 1 the choices of  $(a, b, c)$  leading to the optimal rate and the number of Monte Carlo samples in order to reach a precision  $\delta$ .

	c	a	b	Rate	MC
no bias ( $C_1 = 0$ )	0	$(0, \infty)$	1	$1/n$	$1/\delta^2$
	$[0, 1/2]$	$(-c, +\infty)$	$1 - 2c$	$1/n^{1-c}$	$1/\delta^2$
	$[0, 1]$	$-c$	$(1 - 2c, \infty) \cap [0, \infty)$	$1/n^{1-c}$	$1/\delta^{(1+b)/(1-c)}$
with bias ( $C_1 > 0$ )	0	$(0, \infty)$	1	$1/n$	$1/\delta^2$
	$[0, 1]$	$(0, \infty)$	$1 - c$	$1/n^{1-c}$	$1/\delta^{(2-c)/(1-c)}$
	$[0, 1]$	$-c$	$(1 - c, \infty)$	$1/n^{1-c}$	$1/\delta^{(1+b)/(1-c)}$
$C_1 = C_2 = 0$	$[0, 1]$	$(-1, \infty)$	-	$1/n^{1-c}$	-

TABLE 1. [Averaged Perturbed Proximal Gradient] Values of  $(a, b, c)$  in order to reach the rate of convergence **Rate**. The column MC reports the number of Monte Carlo samples in this strategy to reach a precision  $\mathcal{E}_n = O(\delta)$ . As a reference, the last row reports the rate when  $\eta_n = 0$ .

We now investigate the rates of convergence of the Perturbed FISTA algorithm with Monte Carlo approximation that can be deduced from Corollary 14 as a function of the design parameters  $m_n$  and  $\gamma_n$ . Here again, we assume that there exist  $C_1, C_2 \geq 0$  and  $B > 0$  such that

$$\mathbb{E}[\epsilon_n^{(1)}] \leq C_1 m_{n+1}^{-1}, \quad \mathbb{E}[\epsilon_n^{(2)}] \leq C_2 m_{n+1}^{-1}, \quad \text{and} \quad \mathbb{P} \left( \sup_{n \in \mathbb{N}} \|\theta_n - \theta_\star\| \leq B \right) = 1, \quad (42)$$

and we choose

$$m_n = C_b n^b, \quad \gamma_n = C_c n^{-c},$$

with  $b, c \geq 0$ , and  $C_b, C_c > 0$ .

Let us discuss the rate of convergence that can be deduced from Corollary 14 item ii when  $t_n \sim Cn$  as  $n \rightarrow \infty$ , which is the case for example for the sequence given by (30); note that the rate of decay is at most  $n^c/t_n^2 \sim 1/n^{2-c}$  which implies that hereafter, we choose  $c \in [0, 2)$ .

Consider first the unbiased case ( $C_1 = 0$  in (42)); then the maximal rate of convergence  $n^{c-2}$  can be reached by choosing  $b > 3 - 2c$ . It is therefore possible to reach the maximal rate  $n^{-2}$  (which is the rate of the FISTA algorithm, see (Beck and Teboulle, 2009, Theorem 1.4)) by choosing  $c = 0$  (i.e. a constant step-size) and  $b > 3$  (i.e. a number of Monte Carlo samples which increases as  $n^3$ ). The number of iterations to reach  $\mathbb{E}[F(\theta_n)] - F(\theta_*) \leq \delta$  is  $O(n^{1/(2-c)})$ ; since the number of Monte Carlo after  $n$  iterations is  $O(n^{1+b})$  then,  $O(\delta^{-(1+b)/(2-c)})$  samples are required to reach a precision  $\delta$ . Note that since  $b > 3 - 2c$ , we have  $(1+b)/(2-c) > 2$  that is this computational cost is always larger than the computational cost of the plain stochastic proximal gradient algorithm (see Table 1).

In the biased case (i.e.  $C_1 > 0$  in (42)), the optimal rate  $n^{c-2}$  can be reached by choosing  $b > 3 - c$ . Here again, it is possible to reach the rate  $n^{-2}$  by choosing  $c = 0$  and  $b > 3$ .

We report in Table 2 the conclusions of this discussion.

	c	b	Rate	MC
no bias ( $C_1 = 0$ )	0	$(3, \infty)$	$1/n^2$	$1/\delta^{(b+1)/2}$
	$[0, 2)$	$(3 - 2c, +\infty) \cap [0, \infty)$	$1/n^{2-c}$	$1/\delta^{(b+1)/(2-c)}$
with bias ( $C_1 > 0$ )	0	$(3, \infty)$	$1/n^2$	$1/\delta^{(b+1)/2}$
	$[0, 1)$	$(3 - 2c, \infty)$	$1/n^{2-c}$	$1/\delta^{(b+1)/(2-c)}$
	$[1, 2)$	$(2 - c, \infty)$	$1/n^{2-c}$	$1/\delta^{(b+1)/(2-c)}$
$C_1 = C_2 = 0$	$[0, 2)$	-	$1/n^{2-c}$	-

TABLE 2. [Perturbed FISTA] Values of  $(b, c)$  in order to reach the rate of convergence **Rate**, when  $t_n = O(n^2)$ . The column **MC** reports the number of Monte Carlo samples in this strategy to reach a precision  $\mathbb{E}[F(\theta_n)] - F(\theta_*) = O(\delta)$ . As a reference, the last row reports the rate when  $\tilde{\eta}_n = 0$ .

We conclude this section by a comparison of the approximation of  $F(\theta_*)$  provided by the weighted approach and by the perturbed FISTA. As shown in Table 1 and Table 2, it is possible to choose a step-size sequence and a number of Monte Carlo samples  $m_n$  for these stochastic algorithms to reach the same rate of convergence as the deterministic ones, that is  $n^{-1}$  for the first algorithm and  $n^{-2}$  for the second one. But, taking into account the number of simulations requires to obtain a given precision  $\delta$ , the perturbed averaged Proximal Gradient has a faster rate of convergence than the perturbed FISTA.

## 5. EXAMPLE: HIGH-DIMENSIONAL LOGISTIC REGRESSION WITH RANDOM EFFECTS

We illustrate the results using Example 1. Note that the assumptions H1 and H3 hold but H2 is not in general satisfied. Nevertheless, the numerical study below shows that the conclusions reached in section 3 and section 4 provide useful information to tune the design parameters of the algorithms.

The gradient of the negative log-likelihood for the logistic regression model with random effect is  $-\nabla\ell(\theta) = \int H_\theta(\mathbf{u})\pi_\theta(\mathbf{u})d\mathbf{u}$  with  $\theta = (\beta, \sigma)$ ,  $\pi_\theta$  is given by (5) and

$$H_\theta(\mathbf{u}) = -\sum_{i=1}^N (Y_i - F(x'_i\beta + \sigma z'_i\mathbf{u})) \begin{bmatrix} x_i \\ z'_i\mathbf{u} \end{bmatrix}. \quad (43)$$

The distribution  $\pi_\theta$  is sampled using the MCMC sampler proposed in Polson et al. (2013) based on data-augmentation - see also Choi and Hobert (2013). We write  $-\nabla\ell(\theta) = \int_{\mathbb{R}^q \times \mathbb{R}^N} H_\theta(\mathbf{u})\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) d\mathbf{u}d\mathbf{w}$  where  $\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w})$  is defined for  $\mathbf{u} \in \mathbb{R}^q$  and  $\mathbf{w} = (w_1, \dots, w_N) \in \mathbb{R}^N$  by

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = \left( \prod_{i=1}^N \bar{\pi}_{\text{PG}}(w_i; x'_i\beta + \sigma z'_i\mathbf{u}) \right) \pi_\theta(\mathbf{u});$$

in this expression,  $\bar{\pi}_{\text{PG}}(\cdot; c)$  is the density of the Polya-Gamma distribution on the positive real line with parameter  $c$  given by

$$\bar{\pi}_{\text{PG}}(w; c) = \cosh(c/2) \exp(-wc^2/2) \rho(w) \mathbb{1}_{\mathbb{R}^+}(w),$$

where  $\rho(w) \propto \sum_{k \geq 0} (-1)^k (2k+1) \exp(-(2k+1)^2/(8w)) w^{-3/2}$  (see (Biane et al., 2001, Section 3.1)). Thus, we have

$$\tilde{\pi}_\theta(\mathbf{u}, \mathbf{w}) = C_\theta \phi(\mathbf{u}) \prod_{i=1}^N \exp(\sigma(Y_i - 1/2)z'_i\mathbf{u} - w_i(x'_i\beta + \sigma z'_i\mathbf{u})^2/2) \rho(w_i) \mathbb{1}_{\mathbb{R}^+}(w_i),$$

where  $\ln C_\theta = -N \ln 2 - \ell(\theta) + \sum_{i=1}^N (Y_i - 1/2)x'_i\beta$ . This target distribution can be sampled using a Gibbs algorithm: given the current value  $(\mathbf{u}^t, \mathbf{w}^t)$  of the chain, the next point is obtained by sampling  $\mathbf{u}^{t+1}$  under the conditional distribution of  $\mathbf{u}$  given  $\mathbf{w}^t$ , and  $\mathbf{w}^{t+1}$  under the conditional distribution of  $\mathbf{w}$  given  $\mathbf{u}^{t+1}$ . In the present case, these conditional distributions are given respectively by

$$\tilde{\pi}_\theta(\mathbf{u}|\mathbf{w}) \equiv N_q(\mu_\theta(\mathbf{w}); \Gamma_\theta(\mathbf{w})) \quad \tilde{\pi}_\theta(\mathbf{w}|\mathbf{u}) = \prod_{i=1}^N \bar{\pi}_{\text{PG}}(w_i; |x'_i\beta + \sigma z'_i\mathbf{u}|)$$

with

$$\Gamma_\theta(\mathbf{w}) = \left( I + \sigma^2 \sum_{i=1}^N w_i z_i z'_i \right)^{-1}, \quad \mu_\theta(\mathbf{w}) = \sigma \Gamma_\theta(\mathbf{w}) \sum_{i=1}^N ((Y_i - 1/2) - w_i x'_i\beta) z_i. \quad (44)$$

Exact samples of these conditional distributions can be obtained (see (Polson et al., 2013, Algorithm 1) for sampling under a Polya-Gamma distribution).

We test the algorithms with  $N = 500$ ,  $p = 1,000$  and  $q = 5$ . We generate the  $N \times p$  covariates matrix  $X$  columnwise, by sampling a stationary  $\mathbb{R}^N$ -valued autoregressive model with parameter  $\rho = 0.8$  and Gaussian noise  $\sqrt{1 - \rho^2} \mathcal{N}_N(0, I)$ . We generate the vector of regressors  $\beta_{\text{true}}$  from the uniform distribution on  $[1, 5]$  and randomly set 98% of the coefficients to zero. The variance of the random effect is set to  $\sigma^2 = 0.1$ . We consider a repeated measurement setting so that  $z_i = e_{[iq/N]}$  where  $\{e_j, j \leq q\}$  is the canonical basis of  $\mathbb{R}^q$  and  $\lceil \cdot \rceil$  denotes the upper integer part. With such a simple expression for the random effect, we will be able to approximate the value  $F(\theta)$  in order to illustrate the theoretical results obtained in this paper. We use the Lasso penalty ( $\alpha = 1$  in (3)) with  $\lambda = 30$ .

We first illustrate the ability of Monte Carlo Proximal Gradient algorithms to find a minimizer of  $F$ . We compare the following algorithms

- (i) the Monte Carlo proximal gradient algorithm with  $\gamma_n = \gamma = 0.005$ ,  $m_n = 200 + n$  (Algo 1),  $\gamma_n = \gamma = 0.001$ ,  $m_n = 200 + n$  (Algo 2) and  $\gamma_n = 0.05/\sqrt{n}$  and  $m_n = 270 + \lceil \sqrt{n} \rceil$  (Algo 3).
- (ii) the Monte Carlo FISTA with  $\gamma_n = \gamma = 0.001$ ,  $m_n = 45 + \lceil n^{3.1}/6000 \rceil$  (Algo F1);  $\gamma_n = 0.005 \wedge (0.1/n)$ ,  $m_n = 155 + \lceil n^{2.1}/100 \rceil$  (Algo F2). In both cases,  $t_n$  is given by (30).

Each algorithm is run for 150 iterations. The batch sizes  $\{m_n, n \geq 0\}$  are chosen so that after 150 iterations, each algorithm used approximately the same number of Monte Carlo samples. We denote by  $\beta_\infty$  the value obtained at iteration 150. A path of the relative error  $\|\beta_n - \beta_\infty\|/\|\beta_\infty\|$  is displayed on Figure 1[right] for each algorithm; a path of the sensitivity  $\text{SEN}_n$  and of the precision  $\text{PRE}_n$  defined by

$$\text{SEN}_n = \frac{\sum_i \mathbb{1}_{\{|\beta_{n,i}| > 0\}} \mathbb{1}_{\{|\beta_{\infty,i}| > 0\}}}{\sum_i \mathbb{1}_{\{|\beta_{\infty,i}| > 0\}}}, \quad \text{PRE}_n = \frac{\sum_i \mathbb{1}_{\{|\beta_{n,i}| > 0\}} \mathbb{1}_{\{|\beta_{\infty,i}| > 0\}}}{\sum_i \mathbb{1}_{\{|\beta_{n,i}| > 0\}}},$$

are displayed on Figure 2. All these sequences are plotted versus the total number of Monte Carlo samples up to iteration  $n$ . These plots show that the rate of convergence depends on the step-size sequence  $\{\gamma_n, n \in \mathbb{N}\}$  and the batch-size sequence  $\{m_n, n \in \mathbb{N}\}$ . Fixed stepsize strategies are preferable to decreasing step-size; these findings are consistent with Table 1-biased case. On Figure 1[left], we report on the bottom row the indices  $j$  such that  $\beta_{\text{true},j}$  is non null and on the top row, the indices  $j$  such that  $\beta_{\infty,j}$  given by Algo 1 is non null.

We now study the convergence of  $\{F(\theta_n), n \in \mathbb{N}\}$  where  $\theta_n$  is obtained by one of the algorithms described above. We repeat 50 independent runs for each algorithm and estimate  $\mathbb{E}[F(\theta_n)]$  by the empirical mean over these runs. On Figure 3[left],  $n \mapsto F(\theta_n)$  is displayed for several runs of Algo 1 and Algo 3. The figure shows that all the paths have the same limiting value, which is approximately  $F_\star = 87$ ; we observed the same behavior on the 50 runs of each algorithm. On Figure 3[right], we report the Monte Carlo estimation of  $\mathbb{E}[F(\theta_n)]$  versus the total number of Monte Carlo samples used up to iteration  $n$ . The best strategies are Algo 1 and Algo F1, with an advantage for the Monte Carlo FISTA. Finally, we observe the distribution of  $\{F(\theta_n), n \in \mathbb{N}\}$  through the boxplot of the 50 independent runs. We also consider in this study averaging techniques:

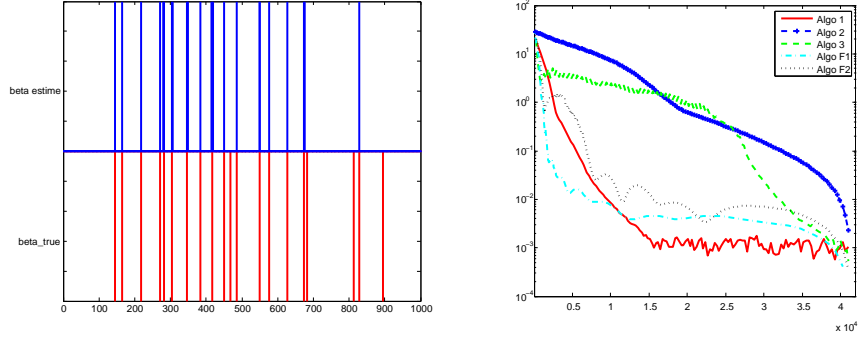


FIGURE 1. [left] The support of the sparse vector  $\beta_\infty$  obtained by Algo 1 (resp.  $\beta_{\text{true}}$ ) on the top (resp. on the bottom). [right] Relative error along one path of each algorithm as a function of the total number of Monte Carlo samples drawn from the initialization of the algorithm.

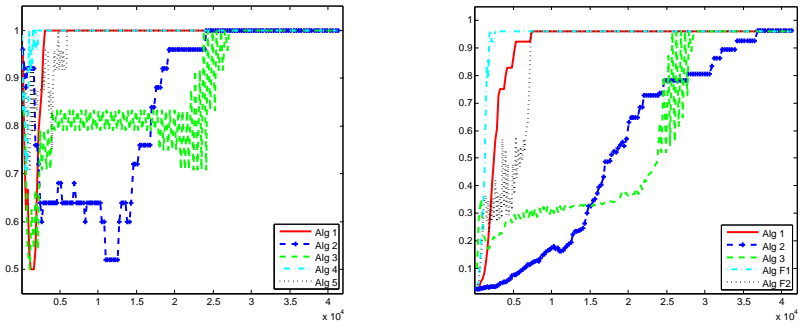


FIGURE 2. The sensitivity  $\text{SEN}_n$  [left] and the precision  $\text{PRE}_n$  [right] along a path, versus the total number of Monte Carlo samples up to time  $n$

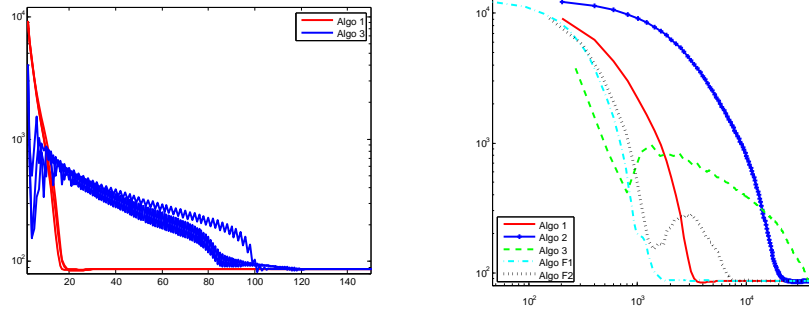


FIGURE 3. [left]  $n \mapsto F(\theta_n)$  for several independent runs. [right]  $\mathbb{E}[F(\theta_n)]$  versus the total number of Monte Carlo samples up to iteration  $n$

- (iii) (Algo W)  $\bar{\theta}_n$  is the weighted average of the output of Algo 1 (see (28)) with three different strategies for the averaging sequence: decreasing weight  $a_n = 1/n^{0.1}$ , uniform weights  $a_n = 1$  and increasing weights  $a_n = \sqrt{n}$ .

The averaging procedure can be implemented in parallel of Algo 1, and of course is of interest when computed from the outputs in the convergence phase of Algo 1. The averaging process is started at iteration  $n = 35$ , discarding the previous estimates. On Figure 4[left], we compare the three strategies for Algo W. For  $n = 50, 100, 150$ , the boxplots of  $F(\bar{\theta}_n)$  are displayed with - from left to right - the decreasing, the uniform and the increasing weight sequences  $\{a_n, n \in \mathbb{N}\}$ . It appears that an increasing weight sequence is more efficient to reduce the fluctuations of  $F(\bar{\theta}_n)$  around its limiting value. On Figure 4[right], we compare Algo 1, Algo F1 and Algo W with increasing weight sequence after the same number of Monte Carlo samples namely 10500, 24200, 41300; this corresponds to iteration  $n = 50, 100, 150$  for Algo 1 and Algo W and  $n = 98, 128, 150$  for Algo F1. Algo F1 is the best one in terms of rate of convergence but seems to have a higher variability. Not surprisingly, there is a gain in averaging the output of Algo 1 in order to reduce the variability.

## 6. PROOFS

### 6.1. Preliminary lemmas.

**Lemma 16.** *Assume that  $g$  is lower semi-continuous and convex. For  $\theta, \vartheta \in \Theta$  and  $\gamma > 0$*

$$\gamma \{g(\text{Prox}_\gamma(\theta)) - g(\vartheta)\} \leq -\langle \text{Prox}_\gamma(\theta) - \vartheta, \text{Prox}_\gamma(\theta) - \theta \rangle. \quad (45)$$

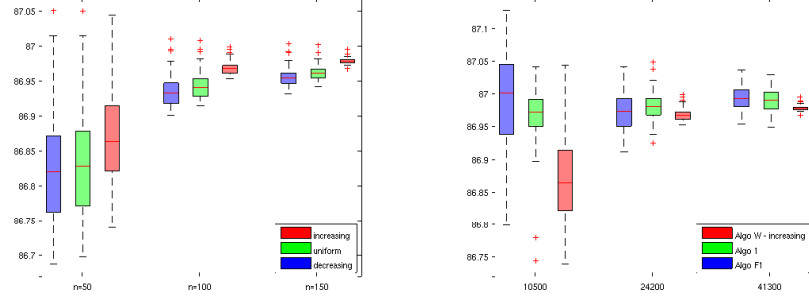


FIGURE 4. [left] Algo W: boxplot of  $F(\bar{\theta}_n)$  for  $n = 50, 100, 150$  with - from left to right - the decreasing, the uniform and the increasing weight sequence. [right] Boxplot of  $F(\theta_n)$  and  $F(\bar{\theta}_n)$  with  $n$  chosen such that the total number of Monte Carlo samples up to time  $n$  is about 10 500, 24 200, 41 300.

For any  $\gamma > 0$ , the operator  $\theta \mapsto \text{Prox}_\gamma(\theta)$  is firmly non-expansive, i.e. for any  $\theta, \vartheta \in \Theta$ ,

$$\langle \text{Prox}_\gamma(\theta) - \text{Prox}_\gamma(\vartheta), \theta - \vartheta \rangle \geq \|\text{Prox}_\gamma(\theta) - \text{Prox}_\gamma(\vartheta)\|^2. \quad (46)$$

In particular, the maps  $\theta \mapsto \text{Prox}_\gamma(\theta)$  and  $\theta \mapsto \theta - \text{Prox}_\gamma(\theta)$  are Lipschitz with Lipschitz constants that can be taken equal to 1.

*Proof.* See (Bauschke and Combettes, 2011, Propositions 4.2., 12.26 and 12.27).  $\square$

Lemma 16 has the following useful consequence.

**Lemma 17.** Assume H1 and let  $\gamma \in (0, 1/L]$ . Then for all  $\theta, \vartheta \in \Theta$ ,

$$-2\gamma (F(\text{Prox}_\gamma(\theta)) - F(\vartheta)) \geq \|\text{Prox}_\gamma(\theta) - \vartheta\|^2 + 2 \langle \text{Prox}_\gamma(\theta) - \vartheta, \vartheta - \gamma \nabla f(\vartheta) - \theta \rangle. \quad (47)$$

If in addition  $f$  is convex, then for all  $\theta, \vartheta, \xi \in \Theta$ ,

$$-2\gamma (F(\text{Prox}_\gamma(\theta)) - F(\vartheta)) \geq \|\text{Prox}_\gamma(\theta) - \vartheta\|^2 + 2 \langle \text{Prox}_\gamma(\theta) - \vartheta, \xi - \gamma \nabla f(\xi) - \theta \rangle - \|\vartheta - \xi\|^2. \quad (48)$$

*Proof.* Since  $\nabla f$  is Lipschitz, the descent lemma implies that for any  $\gamma^{-1} \geq L$

$$f(p) - f(\vartheta) \leq \langle \nabla f(\vartheta), p - \vartheta \rangle + \frac{1}{2\gamma} \|p - \vartheta\|^2. \quad (49)$$

This inequality applied with  $p = \text{Prox}_\gamma(\theta)$  combined with (45) yields (47). When  $f$  is convex,  $f(\xi) + \langle \nabla f(\xi), \vartheta - \xi \rangle - f(\vartheta) \leq 0$  which, combined again with (45) and (49) applied with  $(p, \vartheta) \leftarrow (\text{Prox}_\gamma(\theta), \xi)$  yields the result.  $\square$

**Lemma 18.** *Assume H1. Then for any  $\gamma > 0$ ,*

$$\|\theta - \gamma \nabla f(\theta) - \vartheta + \gamma \nabla f(\vartheta)\| \leq (1 + \gamma L) \|\theta - \vartheta\|, \quad (50)$$

$$\|T_\gamma(\theta) - T_\gamma(\vartheta)\| \leq (1 + \gamma L) \|\theta - \vartheta\|. \quad (51)$$

*If in addition  $f$  is convex then for any  $\gamma \in (0, 2/L]$ ,*

$$\|\theta - \gamma \nabla f(\theta) - \vartheta + \gamma \nabla f(\vartheta)\| \leq \|\theta - \vartheta\|, \quad (52)$$

$$\|T_\gamma(\theta) - T_\gamma(\vartheta)\| \leq \|\theta - \vartheta\|. \quad (53)$$

*Proof.* (51) and (53) follows from (50) and (52) respectively by the Lipschitz property of the proximal map  $\text{Prox}_\gamma$  (see Lemma 16). (50) follows directly from the Lipschitz property of  $f$  assumed in H1. It remains to prove (52). Since  $f$  is a convex function with Lipschitz-continuous gradients, (Nesterov, 2004, Theorem 2.1.5) shows that, for all  $\theta, \vartheta \in \Theta$ ,  $\langle \nabla f(\theta) - \nabla f(\vartheta), \theta - \vartheta \rangle \geq \frac{1}{L} \|\nabla f(\theta) - \nabla f(\vartheta)\|^2$ . The result follows.  $\square$

**Lemma 19.** *Assume H1. Set  $S_\gamma(\theta) \stackrel{\text{def}}{=} \text{Prox}_\gamma(\theta - \gamma H)$  and  $\eta \stackrel{\text{def}}{=} H - \nabla f(\theta)$ . For any  $\theta \in \Theta$  and  $\gamma > 0$ ,*

$$\|T_\gamma(\theta) - S_\gamma(\theta)\| \leq \gamma \|\eta\|. \quad (54)$$

*For any  $\theta, H \in \Theta$  and  $\gamma \in (0, 1/L]$ ,*

$$|F[S_\gamma(\theta)] - F[T_\gamma(\theta)]| \leq (1 + c\gamma L) \|\eta\| (\gamma \|\eta\| + (1 + c\gamma L) \|\theta - T_\gamma(\theta)\|), \quad (55)$$

*where  $c = 0$  if  $f$  is convex or  $c = 1$  otherwise.*

*Proof.* We have  $\|T_\gamma(\theta) - S_\gamma(\theta)\| = \|\text{Prox}_\gamma(\theta - \gamma \nabla f(\theta)) - \text{Prox}_\gamma(\theta - \gamma H)\|$  and (54) follows since  $\theta \mapsto \text{Prox}_\gamma(\theta)$  is a Lipschitz contraction (see Lemma 16).

Apply (47) with  $\vartheta \leftarrow T_\gamma(\theta)$  and  $\theta \leftarrow \theta - \gamma H$ :

$$\begin{aligned} F(S_\gamma(\theta)) - F(T_\gamma(\theta)) &\leq \frac{1}{\gamma} \langle T_\gamma(\theta) - S_\gamma(\theta), T_\gamma(\theta) - \gamma \nabla f(T_\gamma(\theta)) - \theta + \gamma H \rangle \\ &\leq \frac{1}{\gamma} \|T_\gamma(\theta) - S_\gamma(\theta)\| (\|T_\gamma(\theta) - \gamma \nabla f(T_\gamma(\theta)) - \theta + \gamma \nabla f(\theta)\| + \gamma \|\eta\|). \end{aligned} \quad (56)$$

By plugging (54) and (50,52) in (56), we get

$$F(S_\gamma(\theta)) - F(T_\gamma(\theta)) \leq \|\eta\| (\gamma \|\eta\| + (1 + c\gamma L) \|\theta - T_\gamma(\theta)\|).$$

Similarly, we apply the same inequality with  $\vartheta \leftarrow S_\gamma(\theta)$ , and  $\theta \leftarrow \theta - \gamma \nabla f(\theta)$  to get

$$\begin{aligned} F(S_\gamma(\theta)) - F(T_\gamma(\theta)) &\geq \frac{1}{\gamma} \langle T_\gamma(\theta) - S_\gamma(\theta), S_\gamma(\theta) - \gamma \nabla f(S_\gamma(\theta)) - \theta + \gamma \nabla f(\theta) \rangle \\ &\geq -(1 + c\gamma L) \|\eta\| \|\theta - S_\gamma(\theta)\| \\ &\geq -(1 + c\gamma L) \|\eta\| (\|\theta - T_\gamma(\theta)\| + \gamma \|\eta\|), \end{aligned}$$

where we used again (54) and Lemma 18. The results follows.  $\square$



**Lemma 20.** *Assume H1. For any  $\gamma > 0$*

$$\{\theta : \theta = T_\gamma(\theta)\} = \{\theta \in \text{Dom}(g) : 0 \in \nabla f(\theta) + \partial g(\theta)\} . \quad (57)$$

*For any  $\gamma \in (0, 1/L]$  and  $\theta \in \Theta$ ,  $F(T_\gamma(\theta)) - F(\theta) \leq 0$  and  $F \circ T_\gamma(\theta) = F(\theta)$  iff  $\theta = T_\gamma(\theta)$ . If in addition  $f$  is convex, then*

$$F(T_\gamma(\theta)) - F(\theta) \leq -\frac{1}{2\gamma} \|T_\gamma(\theta) - \theta\|^2 . \quad (58)$$

*Proof.* Let  $\gamma > 0$ . By (14), we have

$$\theta = T_\gamma(\theta) \iff \theta = \text{argmin}_\vartheta Q_\gamma(\vartheta|\theta) \iff 0 \in \partial_u Q_\gamma(u|\theta)|_{u=\theta} \text{ and } |g(\theta)| < \infty ,$$

where  $\partial_u$  denotes the sub-differential of the convex function  $u \mapsto Q_\gamma(u|\theta)$ . The proof of (57) is concluded upon noting that  $\partial_u Q_\gamma(u|\theta) = \nabla f(\theta) + \partial g(\theta)$ .

Let  $\gamma \in (0, 1/L]$ . We have from the MM approach (see section 2), for any  $\theta, \vartheta$  and  $\gamma \leq 1/L$ :  $F(\vartheta) \leq Q_\gamma(\vartheta|\theta)$ ,  $F(\theta) = Q_\gamma(\theta|\theta)$  and by definition of  $T_\gamma(\theta)$ ,  $Q_\gamma(T_\gamma(\theta)|\theta) \leq Q_\gamma(\vartheta|\theta)$ . Therefore

$$F(T_\gamma(\theta)) \leq Q_\gamma(T_\gamma(\theta)|\theta) \leq Q_\gamma(\theta|\theta) = F(\theta) .$$

The above inequality also implies that if  $F \circ T_\gamma(\theta) = F(\theta)$  then  $Q_\gamma(T_\gamma(\theta)|\theta) = Q_\gamma(\theta|\theta)$  which implies that  $T_\gamma(\theta) = \theta$ .

When  $f$  is convex, the inequality (58) is a consequence of (48) applied with  $\vartheta \leftarrow \theta$ , and  $\theta \leftarrow \theta - \gamma \nabla f(\theta)$ .  $\square$

**Lemma 21.** *Let  $\{u_n, n \in \mathbb{N}\}$ ,  $\{v_n, n \in \mathbb{N}\}$  and  $\{e_n, n \in \mathbb{N}\}$  be sequences satisfying  $u_n^2 \leq v_n + \sum_{k=0}^n u_k e_k$  and  $2v_n + \sum_{k=0}^n e_k^2 \geq 0$ . Then for any  $n \geq 0$ ,*

$$\sup_{0 \leq k \leq n} \left| u_k - \frac{e_k}{2} \right| \leq \mathcal{U} \left( v_n + \frac{1}{2} \sum_{k=0}^n e_k^2, \frac{1}{2} \sum_{k=0}^{n-1} |e_k| \right)$$

*with the convention that  $\sum_{k=0}^{-1} = 0$  and  $\mathcal{U}(a, b) \stackrel{\text{def}}{=} b + \sqrt{a + b^2}$ .*

*Proof.* The proof is adapted from (Schmidt et al., 2011, Lemma 1). For any  $n \geq 1$ ,

$$\left( u_n - \frac{e_n}{2} \right)^2 \leq v_n + \frac{1}{4} e_n^2 + \sum_{k=0}^{n-1} u_k e_k \leq v_n + \frac{1}{2} \sum_{k=0}^n e_k^2 + \sum_{k=0}^{n-1} \left( u_k - \frac{e_k}{2} \right) e_k .$$

Set

$$A_n \stackrel{\text{def}}{=} v_n + \frac{1}{2} \sum_{k=0}^n e_k^2 \quad B_n \stackrel{\text{def}}{=} \frac{1}{2} \sum_{k=0}^n |e_k| \quad s_n \stackrel{\text{def}}{=} \sup_{0 \leq k \leq n} \left| u_k - \frac{e_k}{2} \right| .$$

Then  $s_n^2 \leq s_{n-1}^2 \vee \{A_n + s_{n-1} 2B_{n-1}\}$ . By induction (note that  $s_0 \leq \sqrt{A_0}$  and  $B_{-1} = 0$ ), this yields for any  $n \geq 0$ ,

$$0 \leq s_n \leq B_{n-1} + (B_{n-1}^2 + A_n)^{1/2} .$$

$\square$

**6.2. Proof of Theorem 4.** Since the sequence  $\{\theta_n, n \in \mathbb{N}\}$  is in the compact set  $\mathcal{K}$ , it possesses at least one accumulation point in  $\mathcal{K}$ . For the proof of the other statements, we apply (Meyer, 1976, Theorem 3.1): let us check its assumptions and show that  $T_\gamma$  is a point-to-set mapping, uniformly compact, strictly monotonic and upper semi-continuous (see the definitions (Meyer, 1976, p.109)).

Under H1,  $T_\gamma$  is a point-to-point mapping. Since  $\mathcal{K}$  is compact, there exist  $\theta_0$  and  $r > 0$  such that for any  $\theta \in \mathcal{K}$ ,  $\|\theta - \theta_0\| \leq r$ . By Lemma 18,  $\|T_\gamma(\theta) - T_\gamma(\theta_0)\| \leq r$  thus showing that  $T_\gamma(\mathcal{K})$  is included in a compact set. Hence, we proved that  $T_\gamma$  is uniformly compact.

Let  $\theta \in \mathcal{K}$  and  $\{\theta_n, n \in \mathbb{N}\}$  be a  $\mathcal{K}$ -valued sequence such that  $\lim_n \theta_n = \theta$ . Assume that the sequence  $\{T_\gamma(\theta_n), n \in \mathbb{N}\}$  converges to  $\theta'$ . Let us prove that  $\theta' = T_\gamma(\theta)$ . By Lemma 18,

$$\begin{aligned} \|T_\gamma(\theta) - \theta'\| &\leq \|T_\gamma(\theta) - T_\gamma(\theta_n)\| + \|T_\gamma(\theta_n) - \theta'\| \\ &\leq (1 + \gamma L) \|\theta_n - \theta\| + \|T_\gamma(\theta_n) - \theta'\|. \end{aligned}$$

The RHS tends to zero as  $n \rightarrow \infty$  thus proving that  $\theta' = T_\gamma(\theta)$ . Hence the mapping  $T_\gamma$  is upper-semicontinuous.

Finally, Lemma 20 shows that  $T_\gamma$  is strictly monotonic w.r.t. the function  $F$ .

**6.3. Proof of Corollary 5.** Let us prove that  $\mathcal{L}$  is not empty iff the sequence  $\{\theta_n, n \in \mathbb{N}\}$  is compact. If the sequence is compact, it possesses an accumulation point which is in the set  $\mathcal{L}$  by Theorem 4; hence,  $\mathcal{L}$  is not empty. Conversely, if  $\mathcal{L}$  is not empty, then there exists  $\theta_\star$  such that  $\theta_\star = T_\gamma(\theta_\star)$  which implies by Lemma 18 and a trivial induction

$$\|\theta_{n+1} - \theta_\star\| = \|T_\gamma(\theta_n) - T_\gamma(\theta_\star)\| \leq \|\theta_n - \theta_\star\| \leq \dots \leq \|\theta_0 - \theta_\star\|.$$

Hence  $\theta_{n+1}$  is in the closed ball centered at  $\theta_\star$  with radius  $\|\theta_0 - \theta_\star\|$ .

**6.4. Proof of Theorem 6.** The claims are a consequence of the deterministic result for random iterative maps developed in (Fort and Moulines, 2003, Proposition 9): let us check the assumptions of this Proposition.

Since  $\limsup_n \|\theta_n\| < \infty$ , there exists a compact set  $\mathcal{K}$  such that  $\theta_n \in \mathcal{K}$  for any  $n$ . By Lemma 18,  $\theta \mapsto T_\gamma(\theta)$  is continuous which implies that  $\mathcal{L}$  is closed; hence,  $\mathcal{L} \cap \mathcal{K}$  is compact.

By Lemma 20, for any  $\theta \in \Theta$ ,  $F(T_\gamma(\theta)) - F(\theta) \leq 0$  and for any compact  $\mathcal{K} \subseteq \Theta \setminus \mathcal{L}$ ,  $\inf_{\mathcal{K}} F \circ T_\gamma - F < 0$  (note that since  $F \circ T_\gamma - F$  is lower semi-continuous, it achieves its infimum on any compact set (see e.g. (Bauschke and Combettes, 2011, Theorem 1.28))). Note also that we can assume that  $F$  is positive by replacing  $F$  with  $\exp(F)$ . Hence,  $F$  is a Lyapunov function relative to  $(T_\gamma, \mathcal{L})$  (see (Fort and Moulines, 2003, p. 1243) for the definitions).

We now prove that for  $\theta_\star \in \mathcal{L}$  (which exists since  $\mathcal{L}$  is not empty) and for any  $B > 0$ ,

$$\lim_{n \rightarrow \infty} |F(\theta_{n+1}) - F(T_\gamma(\theta_n))| \mathbb{1}_{\|\theta_n - \theta_\star\| \leq B} = 0. \quad (59)$$

By Lemma 19,

$$|F(\theta_{n+1}) - F(T_\gamma(\theta_n))| \leq (1 + c\gamma L) \|\eta_{n+1}\| (\gamma \|\eta_{n+1}\| + (1 + c\gamma L) \|\theta_n - T_\gamma(\theta_n)\|).$$

Since  $\theta_\star = T_\gamma(\theta_\star)$  and since  $T_\gamma$  is a Lipschitz function (see Lemma 18), we conclude that there exists  $C > 0$  such that for any  $n \geq 1$ ,

$$|F(\theta_{n+1}) - F(T_\gamma(\theta_n))| \mathbb{1}_{\|\theta_n - \theta_\star\| \leq M} \leq C \|\eta_{n+1}\|.$$

Since  $\lim_n \eta_n = 0$ , the proof of (59) is concluded.

Finally, (Fort and Moulines, 2003, Proposition 9) requires  $F$  to be continuous; a careful reading of their proof shows that this property is used to prove that (i)  $F(\mathcal{L} \cap \mathcal{K})$  is compact and (ii) for any  $\alpha$  small enough,  $F^{-1}(\mathcal{O}_\alpha)$  is an open set of  $\Theta$  where  $\mathcal{O}_\alpha$  is the  $\alpha$ -neighborhood of  $F(\mathcal{L} \cap \mathcal{K})$ . Under our assumptions,  $F$  is lower semi-continuous and is continuous on the domain of  $g$ . By (57),  $\mathcal{L}$  is in the domain of  $g$ , so  $F(\mathcal{L} \cap \mathcal{K})$  is compact. In addition,  $F(\mathcal{L} \cap \mathcal{K}) \subset \mathbb{R}$ , so  $\mathcal{O}_\alpha \subset \mathbb{R}$  and this implies that  $F^{-1}(\mathcal{O}_\alpha)$  is in the domain of  $g$ .

**6.5. Proof of Theorem 7.** We preface the proof with a preliminary lemma, which might be seen as a deterministic version of the Robbins-Siegmund Lemma

**Lemma 22.** *Let  $\{v_n, n \in \mathbb{N}\}$  and  $\{\chi_n, n \in \mathbb{N}\}$  be non-negative sequences and  $\{\eta_n, n \in \mathbb{N}\}$  be such that  $\sum_n \eta_n$  exists. If for any  $n \geq 0$ ,  $v_{n+1} \leq v_n - \chi_n + \eta_n$  then  $\sum_n \chi_n < \infty$  and  $\lim_n v_n$  exists.*

*Proof.* Set  $w_n = v_n + \sum_{k \geq n} \eta_k + M$  with  $M \stackrel{\text{def}}{=} -\inf_n \sum_{k \geq n} \eta_k$  so that  $\inf_n w_n \geq 0$ . Then

$$0 \leq w_{n+1} \leq v_n - \chi_n + \eta_n + \sum_{k \geq n+1} \eta_k + M \leq w_n - \chi_n.$$

$\{w_n, n \in \mathbb{N}\}$  is non-negative and non increasing; therefore it converges. Furthermore,  $0 \leq \sum_{k=0}^n \chi_k \leq w_0$  so that  $\sum_n \chi_n < \infty$ . The convergence of  $\{w_n, n \in \mathbb{N}\}$  also implies the convergence of  $\{v_n, n \in \mathbb{N}\}$ . This concludes the proof.  $\square$

Let  $\theta_\star$  be a minimizer of  $F$  and set  $F_\star \stackrel{\text{def}}{=} F(\theta_\star)$ . Since  $F$  is convex, we have by (48) applied with  $\theta \leftarrow \theta_n - \gamma_{n+1}H_{n+1}$ ,  $\xi \leftarrow \theta_n$ ,  $\vartheta \leftarrow \theta_\star$ ,  $\gamma \leftarrow \gamma_{n+1}$

$$\|\theta_{n+1} - \theta_\star\|^2 \leq \|\theta_n - \theta_\star\|^2 - 2\gamma_{n+1} (F(\theta_n) - F_\star) - 2\gamma_{n+1} \langle \theta_{n+1} - \theta_\star, \eta_{n+1} \rangle.$$

We write  $\theta_{n+1} - \theta_\star = \theta_{n+1} - T_{\gamma_{n+1}}(\theta_n) + T_{\gamma_{n+1}}(\theta_n) - \theta_\star$ . By Lemma 19,  $\|\theta_{n+1} - T_{\gamma_{n+1}}(\theta_n)\| \leq \gamma_{n+1} \|\eta_{n+1}\|$  so that,

$$\begin{aligned} \|\theta_{n+1} - \theta_\star\|^2 &\leq \|\theta_n - \theta_\star\|^2 - 2\gamma_{n+1} (F(\theta_n) - F_\star) \\ &\quad + 2\gamma_{n+1}^2 \|\eta_{n+1}\|^2 - 2\gamma_{n+1} \langle T_{\gamma_{n+1}}(\theta_n) - \theta_\star, \eta_{n+1} \rangle. \end{aligned} \quad (60)$$

*Proof of (i)* Since  $F(\theta) - F_\star \geq 0$ , we have by iterating (60)

$$\|\theta_{n+1} - \theta_\star\|^2 \leq \|\theta_m - \theta_\star\|^2 + 2 \sum_{k=m}^n \gamma_{k+1}^2 \|\eta_{k+1}\|^2 - 2 \sum_{k=m}^n \gamma_{k+1} \langle T_{\gamma_{k+1}}(\theta_k) - \theta_\star, \eta_{k+1} \rangle.$$

*Proof of (ii)* Under (23), (60) and Lemma 22 show that:  $\lim_n \|\theta_n - \theta_\star\|$  exists and  $\sum_n \gamma_{n+1} \{F(\theta_n) - F_\star\} < \infty$ .

Since  $\sum_n \gamma_n = +\infty$ , there exists a subsequence  $\{\theta_{\phi_n}, n \in \mathbb{N}\}$  such that  $\lim_n F(\theta_{\phi_n}) = F_\star$ . Since  $\{\theta_n, n \in \mathbb{N}\}$  is bounded, we can assume without loss of generality that

$\{\theta_{\phi_n}, n \in \mathbb{N}\}$  converges; let  $\theta_\infty$  be the limiting value. Since  $F$  is convex, it is continuous on its domain and we have  $F(\theta_\infty) = F_\star$ . Hence,  $\theta_\infty \in \mathcal{L}$ . By (22), for any  $m$  and  $n \geq \phi_m$

$$\|\theta_{n+1} - \theta_\infty\|^2 \leq \|\theta_{\phi_m} - \theta_\infty\|^2 - 2 \sum_{k=\phi_m}^n \gamma_{k+1} \{ \langle T_{\gamma_{k+1}}(\theta_k) - \theta_\infty, \eta_{k+1} \rangle + \gamma_{k+1} \|\eta_{k+1}\|^2 \}.$$

For any  $\epsilon > 0$ , there exists  $m$  such that the RHS is upper bounded by  $\epsilon$ . Hence, for any  $n \geq \phi_m$ ,  $\|\theta_{n+1} - \theta_\infty\|^2 \leq \epsilon$ , which proves the convergence of  $\{\theta_n, n \in \mathbb{N}\}$  to  $\theta_\infty$ .

**6.6. Proof of Corollary 8.** Define

$$\Omega_\star \stackrel{\text{def}}{=} \bigcup_{B \in \mathbb{N}} \left\{ \sum_n \gamma_{n+1} \{ \epsilon_n^{(1)} + \gamma_{n+1} \epsilon_n^{(2)} \} \mathbb{1}_{\|\theta_n\| \leq B} < \infty \right\} \cap \left\{ \limsup_n \|\theta_n\| < \infty \right\}.$$

Note that  $\mathbb{P}(\Omega_\star) = 1$ . Using the conditional Borel-Cantelli lemma (see (Chen, 1978, Theorem 1)), for any  $B > 0$ ,  $\sum_n \gamma_{n+1}^2 \|\eta_{n+1}\|^2 \mathbb{1}_{\|\theta_n\| \leq B} < \infty$  on  $\Omega_\star$  since  $\sum_n \gamma_{n+1}^2 \epsilon_n^{(2)} \mathbb{1}_{\|\theta_n\| \leq B} < \infty$  on  $\Omega_\star$ .

Consider now the first term in (23). Set  $\xi_k \stackrel{\text{def}}{=} \gamma_k \langle T_{\gamma_k}(\theta_{k-1}) - \theta_\star, \eta_k \rangle$ . Here again, we prove that for any  $B > 0$ ,  $\sum_k \xi_k \mathbb{1}_{\|\theta_{k-1}\| \leq B}$  converges a.s. Fix  $B > 0$ :

$$\sum_{k \geq 0} |\mathbb{E}[\xi_{k+1} | \mathcal{F}_k]| \mathbb{1}_{\|\theta_k\| \leq B} \leq (B + \|\theta_\star\|) \sum_{k \geq 0} \gamma_{k+1} \epsilon_k^{(1)} \mathbb{1}_{\|\theta_k\| \leq B},$$

where we used that  $\theta_\star = T_\gamma(\theta_\star)$  by definition of  $\mathcal{L}$ , and  $\|T_\gamma(\theta) - T_\gamma(\theta_\star)\| \leq \|\theta - \theta_\star\| \leq \|\theta\| + \|\theta_\star\|$  (see Lemma 18). Since the RHS is finite on  $\Omega_\star$ , then  $\sum_k |\mathbb{E}[\xi_{k+1} | \mathcal{F}_k]| \mathbb{1}_{\|\theta_k\| \leq B} < \infty$  on  $\Omega_\star$ . Define  $M_n \stackrel{\text{def}}{=} \sum_{k=1}^n (\xi_k - \mathbb{E}[\xi_k | \mathcal{F}_{k-1}]) \mathbb{1}_{\|\theta_{k-1}\| \leq B}$ ; it is a martingale w.r.t. the filtration  $\{\mathcal{F}_n, n \in \mathbb{N}\}$ . We have

$$\begin{aligned} \sum_{k \geq 1} \mathbb{E} [ |\xi_k - \mathbb{E}[\xi_k | \mathcal{F}_{k-1}]|^2 | \mathcal{F}_{k-1} ] \mathbb{1}_{\|\theta_{k-1}\| \leq B} &\leq \sum_{k \geq 0} \mathbb{E} [ |\xi_k|^2 | \mathcal{F}_{k-1} ] \mathbb{1}_{\|\theta_{k-1}\| \leq B} \\ &\leq (B + \|\theta_\star\|)^2 \sum_{k \geq 0} \gamma_k^2 \epsilon_k^{(2)} \mathbb{1}_{\|\theta_{k-1}\| \leq B} \end{aligned}$$

and the RHS is finite on  $\Omega_\star$  under (25). By (Hall and Heyde, 1980, Theorem 2.17),  $\lim_n M_n$  exists on  $\Omega_\star$ . This concludes the proof of the almost-sure convergence of  $\sum_k \xi_k \mathbb{1}_{\|\theta_{k-1}\| \leq B}$ .

**6.7. Proof of Theorem 10.** We first apply (48) with  $\theta \leftarrow \theta_j - \gamma_{j+1} H_{j+1}$ ,  $\xi \leftarrow \theta_j$ ,  $\vartheta \leftarrow \theta_\star$ ,  $\gamma \leftarrow \gamma_{j+1}$ :

$$V(\theta_{j+1}) \leq \frac{1}{2\gamma_{j+1}} (\|\theta_j - \theta_\star\|^2 - \|\theta_{j+1} - \theta_\star\|^2) - \langle \theta_{j+1} - \theta_\star, \eta_{j+1} \rangle,$$

since  $V(\theta_\star) = 0$ . Multiplying both side by  $a_{j+1}$  gives:

$$\begin{aligned} a_{j+1}V(\theta_{j+1}) &\leq \frac{1}{2} \left( \frac{a_{j+1}}{\gamma_{j+1}} - \frac{a_j}{\gamma_j} \right) \|\theta_j - \theta_\star\|^2 + \frac{a_j}{2\gamma_j} \|\theta_j - \theta_\star\|^2 \\ &\quad - \frac{a_{j+1}}{2\gamma_{j+1}} \|\theta_{j+1} - \theta_\star\|^2 - a_{j+1} \langle \theta_{j+1} - \theta_\star, \eta_{j+1} \rangle. \end{aligned}$$

Summing from  $j = 0$  to  $n - 1$  gives

$$\begin{aligned} \frac{a_n}{2\gamma_n} \|\theta_n - \theta_\star\|^2 + \sum_{j=1}^n a_j V(\theta_j) &\leq \frac{1}{2} \sum_{j=1}^n \left( \frac{a_j}{\gamma_j} - \frac{a_{j-1}}{\gamma_{j-1}} \right) \|\theta_{j-1} - \theta_\star\|^2 \\ &\quad - \sum_{j=1}^n a_j \langle \theta_j - \theta_\star, \eta_j \rangle + \frac{a_0}{2\gamma_0} \|\theta_0 - \theta_\star\|^2. \quad (61) \end{aligned}$$

We decompose  $\langle \theta_j - \theta_\star, \eta_j \rangle$  as follows:

$$\langle \theta_j - \theta_\star, \eta_j \rangle = \langle \theta_j - T_{\gamma_j}(\theta_{j-1}), \eta_j \rangle + \langle T_{\gamma_j}(\theta_{j-1}) - \theta_\star, \eta_j \rangle.$$

By Lemma 19, we get  $|\langle \theta_j - T_{\gamma_j}(\theta_{j-1}), \eta_j \rangle| \leq \gamma_j \|\eta_j\|^2$  which concludes the proof.

**6.8. Proof of Corollary 11.** By conditioning on  $\mathcal{F}_{j-1}$  and using that  $T_{\gamma_j}(\theta_\star) = \theta_\star$  and  $T_{\gamma_j}$  is a 1-Lipschitz operator (see Lemma 18), we have

$$\|\mathbb{E}[\langle T_{\gamma_j}(\theta_{j-1}) - \theta_\star, \eta_j \rangle \mid \mathcal{F}_{j-1}]\| \leq \|\theta_{j-1} - \theta_\star\| \|\mathbb{E}[\eta_j \mid \mathcal{F}_{j-1}]\| = \|\theta_{j-1} - \theta_\star\| \epsilon_{j-1}^{(1)}.$$

The corollary now follows from Theorem 10.

**6.9. Proof of Theorem 13.** Set for any  $j \geq 0$ ,

$$\Delta_{j+1} \stackrel{\text{def}}{=} t_j \theta_{j+1} - (t_j - 1) \theta_j = t_j (\theta_{j+1} - \theta_j) + \theta_j. \quad (62)$$

Note that  $\Delta_{j+1} - \bar{\Delta}_j = t_j(\theta_{j+1} - T_{\gamma_{j+1}}(\theta_j))$  and by Lemma 19,

$$\|\Delta_{j+1} - \theta_\star\| - \|\bar{\Delta}_j - \theta_\star\| \leq \gamma_{j+1} t_j \|\check{\eta}_{j+1}\|. \quad (63)$$

**Lemma 23.** Assume H1 and H2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 3, with  $\{t_n, n \in \mathbb{N}\}$  and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive sequences such that  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ . For any minimizer  $\theta_\star \in \mathcal{L}$ , any  $j \geq 1$ ,

$$\begin{aligned} &(\gamma_j t_{j-1}^2 - \gamma_{j+1} t_j (t_j - 1)) V(\theta_j) + \gamma_{j+1} t_j^2 V(\theta_{j+1}) + \frac{1}{2} \|\Delta_{j+1} - \theta_\star\|^2 \\ &\leq \gamma_j t_{j-1}^2 V(\theta_j) + \frac{1}{2} \|\Delta_j - \theta_\star\|^2 - \gamma_{j+1} t_j \langle \Delta_{j+1} - \theta_\star, \check{\eta}_{j+1} \rangle \quad (64) \end{aligned}$$

$$\leq \gamma_j t_{j-1}^2 V(\theta_j) + \frac{1}{2} \|\Delta_j - \theta_\star\|^2 + \gamma_{j+1}^2 t_j^2 \|\check{\eta}_{j+1}\|^2 - \gamma_{j+1} t_j \langle \bar{\Delta}_j - \theta_\star, \check{\eta}_{j+1} \rangle \quad (65)$$

where  $V(\theta) \stackrel{\text{def}}{=} F(\theta) - F(\theta_\star)$ .

*Proof.* Let  $j \geq 1$ . We first apply (48) with  $\vartheta \leftarrow \theta_j$ ,  $\xi \leftarrow \vartheta_j$ ,  $\theta \leftarrow \vartheta_j - \gamma_{j+1}H_{j+1}$  and  $\gamma \leftarrow \gamma_{j+1}$  to get

$$2\gamma_{j+1}V(\theta_{j+1}) \leq 2\gamma_{j+1}V(\theta_j) + \|\theta_j - \vartheta_j\|^2 - \|\theta_{j+1} - \theta_j\|^2 - 2\gamma_{j+1} \langle \theta_{j+1} - \theta_j, \check{\eta}_{j+1} \rangle .$$

We apply again (48) with  $\vartheta \leftarrow \theta_*$  to get

$$2\gamma_{j+1}V(\theta_{j+1}) \leq \|\theta_* - \vartheta_j\|^2 - \|\theta_{j+1} - \theta_*\|^2 - 2\gamma_{j+1} \langle \theta_{j+1} - \theta_*, \check{\eta}_{j+1} \rangle .$$

We now compute a combination of these two inequalities with coefficients  $t_j(t_j - 1)$  and  $t_j$ . This yields

$$\begin{aligned} & 2\gamma_{j+1}t_j^2V(\theta_{j+1}) + t_j(t_j - 1)\|\theta_{j+1} - \theta_j\|^2 + t_j\|\theta_{j+1} - \theta_*\|^2 \\ & \leq 2t_j(t_j - 1)\gamma_{j+1}V(\theta_j) + t_j(t_j - 1)\|\theta_j - \vartheta_j\|^2 + t_j\|\vartheta_j - \theta_*\|^2 \\ & \quad - 2\gamma_{j+1}t_j \langle \Delta_{j+1} - \theta_*, \check{\eta}_{j+1} \rangle . \end{aligned}$$

Then, by using the definition of  $\vartheta_j$ , we obtain

$$\begin{aligned} 2\gamma_{j+1}t_j^2V(\theta_{j+1}) + \|\Delta_{j+1} - \theta_*\|^2 & \leq 2\gamma_j t_{j-1}^2 V(\theta_j) + \|\Delta_j - \theta_*\|^2 - 2\gamma_{j+1}t_j \langle \Delta_{j+1} - \theta_*, \check{\eta}_{j+1} \rangle \\ & \quad - 2(\gamma_j t_{j-1}^2 - \gamma_{j+1}t_j(t_j - 1))V(\theta_j) . \end{aligned}$$

This concludes the proof.  $\square$

*Proof of Theorem 13.* First note that under (29),  $\gamma_j t_{j-1}^2 - \gamma_{j+1}t_j(t_j - 1) \geq 0$ .

(i) Lemma 22 and (65) show that  $\sup_n \gamma_{n+1}t_n^2 V(\theta_{n+1}) < \infty$ . Since  $\lim_n \gamma_{n+1}t_n^2 = +\infty$ , this implies that  $\lim_n V(\theta_n) = 0$  from which we deduce that the limit points of  $\{\theta_n, n \in \mathbb{N}\}$  are in  $\mathcal{L}$ . (ii) is a trivial consequence of Lemma 23. (iii) By iterating (64), we have for any  $n \geq 1$

$$\frac{1}{2}\|\Delta_{n+1} - \theta_*\|^2 \leq \gamma_1 V(\theta_1) + \frac{1}{2}\|\Delta_1 - \theta_*\|^2 + \sum_{k=1}^n \gamma_{k+1}t_k \|\Delta_{k+1} - \theta_*\| \|\check{\eta}_{k+1}\| .$$

This inequality implies that  $\sup_n \|\Delta_n - \theta_*\| < \infty$  (see Lemma 21). By (63) and the assumption  $\sum_n \gamma_{n+1}^2 t_n^2 \|\check{\eta}_{n+1}\|^2 < \infty$ , this yields  $\sup_n \|\bar{\Delta}_n\| < \infty$ .

**6.10. Proof of Corollary 14 and Corollary 15.** Corollary 14 and Corollary 15 are a consequence of Theorem 13 and the following lemma.

For  $B > 0$ , define the stopping-time

$$\tau_B \stackrel{\text{def}}{=} \inf\{n \geq 1, \|\vartheta_n\| > B\} .$$

**Lemma 24.** Assume H1 and H2. Let  $\{\theta_n, n \in \mathbb{N}\}$  be given by Algorithm 3, with  $\{t_n, n \in \mathbb{N}\}$  and  $\{\gamma_n, n \in \mathbb{N}\}$  be positive sequences satisfying (29) and  $\gamma_n \in (0, 1/L]$  for any  $n \geq 1$ .

(i) If  $\sum_n \gamma_{n+1}t_n \{\|\check{\epsilon}_n^{(1)}\| \mathbb{1}_{n < \tau_B} + \gamma_{n+1}t_n \mathbb{E}[\|\check{\epsilon}_n^{(2)}\| \mathbb{1}_{n < \tau_B}]\} < \infty$ , then for any minimizer  $\theta_* \in \mathcal{L}$ ,  $\sup_n \|(\bar{\Delta}_n - \theta_*) \mathbb{1}_{n < \tau_B}\|_2 < \infty$ . In addition

$$\sup_{k \leq n} \|(\bar{\Delta}_k - \theta_*) \mathbb{1}_{k < \tau_B}\|_2 \leq \sup_{k \leq n} \gamma_{k+1}t_k \|\check{\epsilon}_k^{(1)}\| \mathbb{1}_{k < \tau_B} + \mathcal{U}(A_n, B_n)$$

with  $\mathcal{U}(a, b)$  given by Lemma 21 and

$$\begin{aligned} A_n &\stackrel{\text{def}}{=} \gamma_1 \mathbb{E}[V(\theta_1)] + \frac{1}{2} \mathbb{E}[\|\theta_1 - \theta_\star\|^2] + \sum_{k=1}^n \gamma_{k+1}^2 t_k^2 \mathbb{E}[\check{\epsilon}_k^{(2)} \mathbb{1}_{k < \tau_B}] \\ &\quad + \frac{1}{2} \gamma_{n+1}^2 t_n^2 \mathbb{E}[\check{\epsilon}_n^{(2)} \mathbb{1}_{n < \tau_B}] + 2 \sum_{k=0}^n \gamma_{k+1}^2 t_k^2 \|\check{\epsilon}_k^{(1)} \mathbb{1}_{k < \tau_B}\|_2^2 \\ B_n &\stackrel{\text{def}}{=} \sum_{k=0}^{n-1} \gamma_{k+1} t_k \|\check{\epsilon}_k^{(1)} \mathbb{1}_{k < \tau_B}\|_2. \end{aligned}$$

(ii) Assume  $\limsup_n \|\vartheta_n\| < \infty$  a.s. and for any  $B > 0$  there exist positive constants  $\{M_n, n \in \mathbb{N}\}$  satisfying (35). Then  $\sup_n \gamma_{n+1} t_n^2 V(\theta_n) < \infty$  a.s. and for any minimizer  $\theta_\star \in \mathcal{L}$ ,  $\sup_n \|\bar{\Delta}_n - \theta_\star\| < \infty$  a.s. In addition,

$$\sum_n \gamma_{n+1}^2 t_n^2 \|\check{\eta}_{n+1}\|^2 < \infty, \quad \sum_n \gamma_{n+1} t_n \langle \bar{\Delta}_n - \theta_\star, \check{\eta}_{n+1} \rangle \text{ exists a.s.} \quad (66)$$

*Proof.* (i) Let  $\theta_\star \in \mathcal{L}$ . Iterating (65) yields for any  $n \geq 1$ ,

$$\frac{1}{2} \|\Delta_{n+1} - \theta_\star\|^2 \leq \gamma_1 V(\theta_1) + \frac{1}{2} \|\Delta_1 - \theta_\star\|^2 + \sum_{j=1}^n \gamma_{j+1}^2 t_j^2 \|\check{\eta}_{j+1}\|^2 - \sum_{j=1}^n \gamma_{j+1} t_j \langle \bar{\Delta}_j - \theta_\star, \check{\eta}_{j+1} \rangle.$$

By (63), (65) and the inequality  $(a+b)^2 \leq 2a^2 + 2b^2$ , we have

$$\begin{aligned} \frac{1}{4} \|\bar{\Delta}_n - \theta_\star\|^2 &\leq \gamma_1 V(\theta_1) + \frac{1}{2} \|\Delta_1 - \theta_\star\|^2 + \sum_{j=1}^n \gamma_{j+1}^2 t_j^2 \|\check{\eta}_{j+1}\|^2 \\ &\quad + \frac{1}{2} \gamma_{n+1}^2 t_n^2 \|\check{\eta}_{n+1}\|^2 - \sum_{j=1}^n \gamma_{j+1} t_j \langle \bar{\Delta}_j - \theta_\star, \check{\eta}_{j+1} \rangle. \end{aligned}$$

Multiplying by  $\mathbb{1}_{n < \tau_B}$ , using the inequality  $\mathbb{1}_{n < \tau_B} \leq \mathbb{1}_{j < \tau_B}$  for any  $j \leq n$  and applying the expectation and the Cauchy-Schwartz inequality and the inequality yield

$$\begin{aligned} \frac{1}{4} \mathbb{E}[\|\bar{\Delta}_n - \theta_\star\|^2 \mathbb{1}_{n < \tau_B}] &\leq \gamma_1 \mathbb{E}[V(\theta_1)] + \frac{1}{2} \mathbb{E}[\|\Delta_1 - \theta_\star\|^2] + \sum_{j=1}^n \gamma_{j+1}^2 t_j^2 \mathbb{E}[\check{\epsilon}_j^{(2)} \mathbb{1}_{j < \tau_B}] \\ &\quad + \frac{1}{2} \gamma_{n+1}^2 t_n^2 \mathbb{E}[\check{\epsilon}_n^{(2)} \mathbb{1}_{n < \tau_B}] + \sum_{j=1}^n \gamma_{j+1} t_j \|\bar{\Delta}_j - \theta_\star\|_{j < \tau_B} \|\check{\epsilon}_j^{(1)} \mathbb{1}_{j < \tau_B}\|_2. \quad (67) \end{aligned}$$

We then conclude by Lemma 21.

(ii) If (66) holds, then the other statements follow by Lemma 22 and (65) that  $\lim_n 2\gamma_{n+1} t_n^2 V(\theta_n) + \|\Delta_n - \theta_\star\|^2$  exists a.s. This implies  $\sup_n \|\Delta_n\| < \infty$  a.s. and  $\sup_n \gamma_{n+1} t_n^2 V(\theta_n) < \infty$  a.s. By (63), this yields  $\sup_n \|\bar{\Delta}_n\| < \infty$  a.s.

Let us prove (66). The assumption  $\sum_n \gamma_{n+1}^2 t_n^2 M_n < \infty$  and the conditional Borel-Cantelli lemma (see e.g. (Chen, 1978, Theorem 1)) imply that for any  $B > 0$ ,  $\sum_n \gamma_{n+1}^2 t_n^2 \|\check{\eta}_{n+1}\|^2 \mathbb{1}_{n < \tau_B} < \infty$  a.s. This implies in turn that  $\sum_n \gamma_{n+1}^2 t_n^2 \|\check{\eta}_{n+1}\|^2 < \infty$

a.s. We now prove that  $\sum_n \gamma_{n+1} t_n \langle \bar{\Delta}_n - \theta_*, \check{\eta}_{n+1} \rangle < \infty$  a.s. for any  $\theta_* \in \mathcal{L}$ . On one hand,

$$\sum_n \gamma_{n+1} t_n |\langle \bar{\Delta}_n - \theta_*, \mathbb{E}[\check{\eta}_{n+1} | \mathcal{F}_n] \rangle| \mathbb{1}_{n < \tau_B} \leq \sum_n \gamma_{n+1} t_n M_n \{ \|\bar{\Delta}_n\| \mathbb{1}_{n < \tau_B} + \|\theta_*\| \}.$$

By (i), the expectation of the RHS is finite which implies that the LHS is finite a.s. On the other hand,  $\{\sum_{j=1}^n \gamma_{j+1} t_j \langle \bar{\Delta}_j - \theta_*, \check{\eta}_{j+1} - \mathbb{E}[\check{\eta}_{j+1} | \mathcal{F}_j] \rangle \mathbb{1}_{j < \tau_B}, n \in \mathbb{N}\}$  is a martingale w.r.t.  $\mathcal{F}_n$  and it converges a.s. : indeed, for any  $B > 0$ ,

$$\begin{aligned} \sum_{j=1}^n \gamma_{j+1}^2 t_j^2 \mathbb{E}[\|\bar{\Delta}_j - \theta_*\|^2 \mathbb{E}[\|\check{\eta}_{j+1} - \mathbb{E}[\check{\eta}_{j+1} | \mathcal{F}_j]\|^2 | \mathcal{F}_j]] \mathbb{1}_{j < \tau_B} \\ \leq \sup_j \|(\bar{\Delta}_j - \theta_*) \mathbb{1}_{A_j}(B)\|_2^2 \sum_{j \geq 1} \gamma_{j+1}^2 t_j^2 M_j \end{aligned}$$

and the RHS is finite by assumptions and (i); the convergence of the martingale follows from (Hall and Heyde, 1980, Theorem 2.10).  $\square$

#### APPENDIX A. EXAMPLE 1

By using the Cauchy-Schwartz inequality, it holds

$$\begin{aligned} \int \exp(\ell_c(\theta | \mathbf{u})) \phi(\mathbf{u}) d\mathbf{u} &\geq \left( \int \exp(0.5 \ell_c(\theta | \mathbf{u})) \phi(\mathbf{u}) d\mathbf{u} \right)^{1/2} \\ \left( \int \exp(\ell_c(\theta | \mathbf{u})) \|u\|^2 \phi(\mathbf{u}) d\mathbf{u} \right)^2 &\leq \left( \int \exp(0.5 \ell_c(\theta | \mathbf{u})) \phi(\mathbf{u}) d\mathbf{u} \right) \left( \int \exp(3 \ell_c(\theta | \mathbf{u})/2) \|u\|^4 \phi(\mathbf{u}) d\mathbf{u} \right) \end{aligned}$$

which implies that

$$\begin{aligned} \int \|u\|^2 \pi_\theta(\mathbf{u}) d\mathbf{u} &= \frac{\int \exp(\ell_c(\theta | \mathbf{u})) \|u\|^2 \phi(\mathbf{u}) d\mathbf{u}}{\int \exp(\ell_c(\theta | v)) \phi(v) dv} \\ &\leq \left( \int \exp(3 \ell_c(\theta | \mathbf{u})/2) \|\mathbf{u}\|^4 \phi(\mathbf{u}) d\mathbf{u} \right)^{1/2} \end{aligned}$$

Since  $\exp(\ell_c(\theta | \mathbf{u})) \leq 1$  (it is the likelihood of i.i.d. Bernoulli variables) and  $\int \|\mathbf{u}\|^4 \phi(\mathbf{u}) d\mathbf{u} = q(2 + q)$ , we have

$$\sup_{\theta \in \Theta} \int \|u\|^2 \pi_\theta(\mathbf{u}) d\mathbf{u} \leq \sqrt{q(2 + q)}.$$

#### APPENDIX B. EXAMPLE 2

For  $\theta, \vartheta \in \Theta$ , the  $(i, j)$ -th entry of the matrix  $\nabla \ell(\theta) - \nabla \ell(\vartheta)$  is given by

$$(\nabla \ell(\theta) - \nabla \ell(\vartheta))_{ij} = \int_{\mathbf{X}^p} \bar{B}_{ij}(x) \pi_\vartheta(dx) - \int_{\mathbf{X}^p} \bar{B}_{ij}(x) \pi_\theta(dx).$$

For  $t \in [0, 1]$  let

$$\pi_t(dz) \stackrel{\text{def}}{=} \exp(\langle \bar{B}(z), t\vartheta + (1-t)\theta \rangle) / \int \exp(\langle \bar{B}(x), t\vartheta + (1-t)\theta \rangle) \mu(dx),$$



defines a probability measure on  $\mathbf{X}^p$ . It is straightforward to check that

$$(\nabla \ell(\theta) - \nabla \ell(\vartheta))_{ij} = \int \bar{B}_{ij}(x) \pi_1(dx) - \int \bar{B}_{ij}(x) \pi_0(dx),$$

and that  $t \mapsto \int \bar{B}_{ij}(x) \pi_t(dx)$  is differentiable with derivative

$$\begin{aligned} \frac{d}{dt} \int \bar{B}_{ij}(x) \pi_t(dx) &= \int \bar{B}_{ij}(x) \left\langle \bar{B}(x) - \int \bar{B}(z) \pi_t(dz), \vartheta - \theta \right\rangle \pi_t(dx), \\ &= \text{Cov}_{\pi_t}(\bar{B}_{ij}(X), \langle \bar{B}(X), \vartheta - \theta \rangle), \end{aligned}$$

where the covariance is taken assuming that  $X \sim \pi_t$ . Hence

$$\begin{aligned} |(\nabla \ell(\theta) - \nabla \ell(\vartheta))_{ij}| &= \left| \int_0^1 dt \text{Cov}_t(\bar{B}_{ij}(X), \langle \bar{B}(X), \vartheta - \theta \rangle) \right| \\ &\leq \text{osc}(\bar{B}_{ij}) \sqrt{\sum_{k \leq l} \text{osc}^2(\bar{B}_{kl})} \|\theta - \vartheta\|_2. \end{aligned}$$

This implies the inequality (9).

**Acknowledgments:** We are grateful to George Michailidis for very helpful discussions. This work is partly supported by NSF grant DMS-1228164.

#### REFERENCES

- BAUSCHKE, H. H. and COMBETTES, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York. With a foreword by Hedy Attouch. URL <http://dx.doi.org/10.1007/978-1-4419-9467-7>
- BECK, A. and TEOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202.
- BECK, A. and TEOULLE, M. (2010). Gradient-based algorithms with applications to signal-recovery problems. In *Convex optimization in signal processing and communications*. Cambridge Univ. Press, Cambridge, 42–88.
- BERTSEKAS, D. (2012). In *Optimization for Machine Learning* (S. Sra, S. Nowozin and S. Wright, eds.). MIT Press, Cambridge, 85–119.
- BIANE, P., PITMAN, J. and YOR, M. (2001). Probability laws related to the Jacobi theta and Riemann zeta functions, and Brownian excursions. *Bull. Amer. Math. Soc. (N.S.)* **38** 435–465 (electronic). URL <http://dx.doi.org/10.1090/S0273-0979-01-00912-0>
- CHEN, L. (1978). A short note on the Conditional Borel-Cantelli Lemma. *Ann. Probab.* **6** 699–700.
- CHOI, H. M. and HOBERT, J. P. (2013). The polygamma gibbs sampler for bayesian logistic regression is uniformly ergodic. *Electronic Journal of Statistics* **7** 2054–2064.
- COMBETTES, P. and PESQUET, J. (2014). Stochastic quasi-fejer block-coordinate fixed point iterations with random sweeping. Tech. rep., arXiv:1404.7536.
- COMBETTES, P. and WAJS, V. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation* **4** 1168–1200.

- COTTER, A., SHAMIR, O., SREBRO, N. and SRIDHARAN, K. (2011). Better mini-batch algorithms via accelerated gradient methods. In *Advances in Neural Information Processing Systems 24* (J. Shawe-taylor, R. Zemel, P. Bartlett, F. Pereira and K. Weinberger, eds.). 1647–1655.  
URL [http://books.nips.cc/papers/files/nips24/NIPS2011\\_0942.pdf](http://books.nips.cc/papers/files/nips24/NIPS2011_0942.pdf)
- DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159.
- FORT, G. and MOULINES, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* **31** 1220–1259.  
URL <http://dx.doi.org/10.1214/aos/1059655912>
- FORT, G., MOULINES, E. and PRIOURET, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.* **39** 3262–3289.  
URL <http://dx.doi.org/10.1214/11-AOS938>
- HALL, P. and HEYDE, C. (1980). *Martingale Limit Theory and its Application*. Academic Press.
- HU, C., PAN, W. and KWOK, J. T. (2009). Accelerated gradient methods for stochastic optimization and online learning. In *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.).
- JENNRICH, R. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *Ann. Math. Statist.* **40** 633–643.
- JUDITSKY, A. and NEMIROVSKI, A. (2012a). First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. In *Oxford Handbook of Innovation* (S. Sra, S. Nowozin and S. Wright, eds.). MIT Press, Boston, 121–146.
- JUDITSKY, A. and NEMIROVSKI, A. (2012b). First-order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem’s structure. In *Oxford Handbook of Innovation* (S. Sra, S. Nowozin and S. Wright, eds.). MIT Press, Boston, 149–181.
- LAN, G. (2012). An optimal method for stochastic composite optimization. *Math. Program.* **133** 365–397.
- MCLACHLAN, G. and KRISHNAN, T. (2008). *The EM algorithms and Extensions*. Wiley-Interscience; 2 edition.
- MEYER, R. (1976). Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *J. Comput. System. Sci.* **12** 108–121.
- NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609.  
URL <http://dx.doi.org/10.1137/070704277>
- NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization, A basic course*. Kluwer Academic Publishers.
- NESTEROV, Y. E. (1983). A method for solving the convex programming problem with convergence rate  $O(1/k^2)$ . *Dokl. Akad. Nauk SSSR* **269** 543–547.
- NITANDA, A. (2014). Stochastic proximal gradient descent with acceleration techniques. NIPS.

- PARIKH, N. and BOYD, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization* **1** 123–231.
- PETROV, V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Clarendon Press.
- POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. *ArXiv e-prints 1205.0310v3*.
- POLYAK, B. and JUDITSKY, A. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855.
- ROBERT, C. and CASELLA, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics, Springer; 2nd edition.
- ROSASCO, L., VILLA, S. and VU, B. (2014). Convergence of a Stochastic Proximal Gradient Algorithm. Tech. rep., arXiv:1403.5075v3.
- SCHIFANO, E., STRAWDERMAN, R. and WELLS, M. (2010). Majorization-Minimization algorithms for nonsmoothly penalized objective functions. *Electron. J. Statist.* **4** 1258–1299.
- SCHMIDT, M., LE ROUX, N. and BACH, F. (2011). Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*. See also the technical report INRIA-00618152.
- SHAO, J. (2003). *Mathematical Statistics*. Springer texts in Statistics.
- XIAO, L. (2010). Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.* **11** 2543–2596.
- XIAO, L. and ZHANG, T. (2014). A Proximal Stochastic Gradient Method with Progressive Variance Reduction. Tech. rep., arXiv:1403.4699.